ANALYSING HOW THE ARABIDOPSIS CIRCADIAN NETWORK RESPONDS TO TEMPERATURE

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy

By

Benjamin John Wareham

September 2013

Declaration

This thesis is the result of my own work, unless otherwise stated, and is based upon the results from experimental and theoretical work performed as a PhD student between October 2010 and September 2013 in the Institute of Integrative Biology within the University of Liverpool.

Neither this thesis nor any part of it has been submitted as part of an application for another degree or qualification at this or any other University of learning.

Ben Wareham
September 2013

Abstract

The circadian clock is an endogenous 24 hours oscillator found within many organisms. It is involved in controlling gene regulation so that different processes are activated at specific times of the day. The clock maintains regular cycles with a period of approximately 24 hours in a range of conditions such as changing environmental temperature. In the past decade, great steps forward have been made in understanding the network of genes that control the circadian clock, and how these are able to maintain their rhythm in a range of conditions. However, existing models are still limited by which conditions they can accurately simulate.

Here it was shown that the topology of the circadian clock in *Arabidopsis thaliana* changes with changing environmental temperature. This was initially investigated using transcriptomic data from ATH1 arrays. This analysis showed that plant buffering to a changing environmental temperature is controlled at a systems level and are not just controlled by a few genes. These changes occur broadly across different biological functions. However, an in-depth analysis suggests temperature responses are primarily regulated by a balancing act between transcription, translation and protein degradation. Further analysis also identified 13 genes important for temperature compensation. This was confirmed using a delayed fluorescence screen to analysis the circadian rhythm it mutants where these genes had their expression knocked out.

Using a large luciferase data set, it was demonstrated how circadian genes were expressed relative to each other. Initial cluster results suggested that whilst several genes were repeatedly clustered together at different temperatures, the clustering of many genes changed with temperature. This data was then used to create networks using network inference software, which mathematically predicts gene relationships. Network inference successfully recreated networks similar to existing models. However the networks produced for each temperature had significant differences.

Table of Contents

List of Abbreviations

BiNGO	Biological Networks Gene Ontology tool
CAB2	CHLOROPHYLL BINDING PROTEIN A/B
CCA1	CIRCADIAN CLOCK ASSOCIATED 1
CCR2	COLD CIRCADIAN RHYTHM AND RNA BINDING
COP1	CONSTITUTIVE PHOTOMORPHOGENIC 1
CRY	CRYPTOCHROME
CSI	Causal Structure Identification
DF	Delayed Fluorescence
ELF	EARLY FLOWERING
FFT	Fast Fourier Transform
FRQ	FREQUENCY
GI	GIGANTEA
GO	Gene Ontology
GUI	Graphical User Interface
LD	Light Dark cycles
LDLL	Light Dark cycles followed by constant Light
LHY	LATE ELONGATED HYPOCOTYL
LED	Light Emitting Diode
LL	constant Light
LUC	firefly LUCIFERASE
LUX	LUX ARRHYTHMO
mRNA	messenger ribonucleic acid
MS	Murashige and Skoog
NASC	Nottingham Arabidopsis Stock Centre
NI	NIGHT INHIBITED
ODE	Ordinary Differential Equations
PCA	Principal Component Analysis
PHY	PHYTOCHROME
PRR	PSEUDO RESPONSE REGULATOR
PSII	PHOTOSYSTEM II
ReTrOS	Reconstructing Transcription Open Software

RNA	Ribonucleic acid
ROBuST	Regulation of Biological Signalling by Temperature
TAIR	The Arabidopsis Information Resource
TIC	TIME FOR COFFEE
TOC1	TIMING OF CAB EXPRESSION 1
VBSSM	Variational Bayesian State Space Modeling
WC	WHITE COLLAR
WT	Wild type
ZTL	ZEITLUPE

List of Figures

Chapter 1 – Introduction

Chapter 3 – Microarray Analysis

Chapter 4 – Clustering

Chapter 5 – VBSSM

Chapter 6 – CSI

Appendix – Supplemental Data

List of Tables

Chapter 3 – Microarray Analysis

Chapter 4 – Clustering

Chapter 6 – CSI

Appendix – Supplemental Data

Acknowledgements

First I would like to thank my PhD supervisor, Anthony Hall, for all his help and hard work throughout this project. I am grateful for the time spent teaching me the core principles and techniques used throughout my time here as well as the discussions that helped to expand my research. I would also like to thank his post-doc researcher, Peter Gould, for help in the lab and generation of the time series data and microarrays used within my analysis. In addition, I would like to thank the other members of the ROBuST project for continued feedback into various aspects of my research, as well as the funding received for this project. I would like to specifically thank David Rand and Andrew Millar from this group for the informative conversations about network inference and modeling. I would also like to thank my secondary supervisor, Natasha Savage, for her help with Matlab programming as well as her expertise in Boolean modeling. Additionally thanks go to several people for help with specific software packages; Chris Penfold for the help understanding how to utilise his CSI code as well help with adapting it to fit my purpose. Nicholas Heard for his aid with SplineCluster. Silvia Liverani for access to her FFT spline clustering software and discussion on the output. As well as Paul Kirk for attempts to co-cluster luciferase data as well as the idea for consensus clustering and discussions that led to improvements of this technique. Thanks also go to Phillip Antczak for access to the high-powered computer needed to run network inference as well as the University of Liverpool for hosting me for the duration of my PhD.

In addition I am grateful to my friends and family who supported me in both this and my previous research projects. Their interest and conversations have allowed me to improve how I communicate my research to a range of audiences.

Chapter 1 – Introduction

1.1 – Circadian Clock

Within organisms there are a wide range of biological functions that oscillate with predictable patterns. These biological rhythms contain rhythmic genes, which work together to control the speed and timing of the biological function. The expression of these genes, both at the mRNA and protein level, increases and decreases with a set period. The length of these periods determine how they are categorised, but can range from very fast oscillations, such as defecation in *C. elegans* (Liu & Thomas 1994), to very slow oscillations, such as annual migration patterns (Rusak & Zucker 1975). Within these different period lengths, circadian rhythms have had the most research due to their impact on the rest of the metabolic processes. A circadian rhythm is so called because its cycle length is approximately 1 day (latin: circa – about, diem – a day).

1.1.1 – Defining Features of Circadian Rhythms

Rhythmic genes that are synchronised to the external day:night cycles are referred to as diurnal (Aschoff 1963). However organisms are subject to more than just changes in external lighting cues. The environment is frequently changing due to variation in parameters including temperature, humidity, atmospheric pressure, rainfall, as well as changes in perceived light. These changes are not always predictable. Despite this, many of these rhythmic processes, as well as the genes that control them, maintain a fixed period despite the varying environment; some persist even if the signals are completely removed.

This entrained, 24-hour oscillator, which persists in a range of environmental condition, provides an endogenous timer that allows organisms to cue many biological functions, including metabolic processes, to occur at specific times. This can be seen across nature, from sleep/wake cycles in mammals (Mills et al. 1974) to leaf movement in plants (McClung 2006). By having these processes controlled by this endogenous 'clock', organisms can ensure that their biology

continues even if an environmental cue is late or missing. For organisms that are unable to control their environment, having a reliable, buffered, timing mechanism ensures that biological processes retain their timing. This has been shown to be of great evolutionary advantage. When the biology of an organism was accurately controlled by a clock that mimicked the environment, fitness was significantly increased (Dodd 2005; Ouyang et al. 1998). This can be seen in examples where the photoperiod (i.e. the length of days and nights) was altered. Plants grown in these altered time periods had reduced growth when compared to plants grown under normal 12 hours light 12 hours dark cycles (Highkin & Hanson 1954). Gene mutations have been discovered that alter the period of the circadian clock in constant (free-running) conditions. These mutants usually have reduced growth compared to wild type when grown in normal environmental cycles. Similarly when mutants with altered endogenous free-running period lengths were grown in daily light cycles that matched their internal clock, they outperformed the wild type. This has been shown with both long and short period cyanobacteria (Ouyang et al. 1998), as well as *Arabidopsis thaliana* (Dodd 2005).

Lots of research has been done to discover just how important the circadian clock is to the regulation of the transcriptome. In plants, *Arabidopsis thaliana* has been frequently used as a model organism. Through microarray analysis, it was originally thought that between 6 and 16% of the transcriptome was regulated by the clock (Harmer 2000; Edwards 2006). As microarrays were designed to specifically test for circadian control, and multiple datasets were used in unison, this value increased to around 36% (Michael 2003). However, when advances that were made in mammalian clock research were applied to microarrays, virtually all genes were suggested as being under circadian control (Ptitsyn 2008). With this amount of dependency on accurate circadian control, the buffering of the clock to environmental perturbations is crucial for plant survival. However, plants are highly sensitive to environmental cues, some of which are often very subtle such as the ratio of red light to far red light. Plants are able to detect dawn/dusk light changes through this lighting ratio, allowing them to adjust their gene rhythms to better match their external environment

(Wenden et al. 2011). Additionally, temperature alone was capable of entraining the clock within Arabidopsis (McClung et al. 2002). This suggests a mechanism within the clock that not only regulates metabolic speeds to maintain period, but also one that can adjust phases.

In summary, the defining features of circadian rhythms are:

- Oscillating expression levels of mRNA, proteins, metabolites etc. with a roughly 24 hour period, maintained even in the absence of stimulus.
- Buffered against non-specific changes in the environment.
- Entrainable to external stimuli such as light or temperature pulses.

1.1.2 – Plant Responses to Temperature

Plants are subject to a range of environmental temperatures throughout the year. The timing of these temperature changes spans from long seasonal variation to more rapid changes experienced throughout the day/night cycle. As with all poikilotherms, plants are unable to self regulate their own temperature to compensate for these fluctuations. Moreover, the sessile nature of plants means they cannot relocate to areas with more desirable temperatures. Thus, it is critical for plants, perhaps more than any other organism, to develop ways to adapt to the variable temperatures.

Different aspects of this adaption to temperature fluctuation have been studied. At low temperatures where the plant is susceptible to freezing a cold acclimation response is induced (Chinnusamy et al. 2007; M. R. Knight & H. Knight 2012). This response is coordinated by a set of transcription factors called the *C-REPEAT BINDING FACTORS* (*CBF*s). Up regulation of theses genes causes the promotion of an additional set of genes, the *COLD-RESPONSIVE* (*COR*) genes. These help prevent the plant from freezing and gain a tolerance for further exposure to freezing temperatures through a post-translational adaptation to *INDUCER OF CBF EXPRESSION1* (*ICE1)* (Chinnusamy et al. 2007). At the other extreme, when high temperatures become a threat to the plant,

mechanisms to enhance thermotolerance are induced (Queitsch et al. n.d.). This thermotolerence mechanism activates a series of heat-shock proteins (HSP's) that help protect the plant from oxidative stress and protein misfolding. Oxidative stress, which is not immediately countered with antioxidants, cause protein damage. This damage leads to protein misfolding and aggregation. HSP's manage this via a chaperone function, repairing mis-folded proteins and preventing the aggregation of denatured proteins. These pathways are well researched with mechanisms known and integrated into plant biology at a higher level. However, between these two extremes lie ambient temperatures. While these temperatures do not induce full stress responses they have a profound effect on the growth and development of the plant.

Multiple pathways have been identified as responding to ambient temperatures. These include the circadian clock, photobiology, development and growth (Gould et al. 2006; Gould et al. 2013; Portolés et al. 2010). These processes naturally occur faster at higher temperatures and require the activation of additional components to retain constant metabolic rates. There are also changes at a molecular level in response to ambient temperature such as chromatin remodelling (Kumar & Wigge 2010). While responding to temperature is important, the plant has also developed remarkable temperature buffering capabilities. A key example of this is the circadian clock. The circadian clock is a central oscillator of the plant, which has been shown to control the expression of a large proportion of the transcriptome (Ptitsyn 2008). Should the clocks timing be altered, many processes would occur at the wrong time. This is highlighted within the photosystem II pathway involved in photosynthesis, where loss of circadian control leads to a reduction in carbon fixed (Queitsch et al. n.d.; Dodd 2005). However, the clock itself may not be completely buffered, with rhythms running slightly faster at higher temperatures and this shorter period has been show to confer a fitness advantage at higher temperatures, altering the phasing of rhythms to dawn in response to higher temperatures (Gould et al. 2006; Kusakina et al. 2013; Gould et al. 2013; Portolés et al. 2010).

1.1.3 – Role of Temperature Compensation

Temperature compensation is the specific mechanism that buffers the circadian clock against temperature. Within *Arabidopsis,* research has identified several core clock genes essential for temperature compensation. These include: *GIGANTEA* (*GI*) – a gene that is expressed predominately in the evening (Locke et al. 2006) – and *PSEUDO RESPONSE REGULATORS 7* and *9* (*PRR7/9*) – genes that are expressed sequentially in the morning (Makino et al. 2002; Nakamichi 2005). However, it is currently unclear whether this temperature compensation occurs through specific molecular mechanisms, such as the genes listed above, that evolved to produce this effect, or whether temperature compensation is a network wide mechanism with the genes identified acting as major hubs of this network. Additionally, recent work has demonstrated that temperature compensation likely occurs through the light entrainment pathways (Gould et al. 2013). There is also mounting support for temperature compensation being driven by global changes in transcription and translation rates (Sidaway-Lee et al. 2013).

With a functioning temperature compensation mechanism, organisms are able to maintain correct timing of their biology relative to external signals. This results in an organism that maintains the same period and amplitude of gene oscillations. By studying how this is achieved in normal environmental conditions, it may be possible to extend the functioning range of this mechanism, and as such allow plants to be grown in different environments. With the global populations rising, the need for greater food supplies constantly increases. However, recent studies have shown that the trend in increasing temperatures coincides with decreasing crop yields (Lobell et al. 2011). The ability for farmers to match the growing demand for food would become more difficult should these trends continue. Discovering a method to counter the reduced yields that accompany increased temperature will be a powerful tool in improving crop yields in the face of climate change. Understanding how temperature compensation functions to buffer circadian genes may provide a method to maintain high crop yields at higher temperatures.

1.1.3.1 – Temperature Compensation in other Organisms

Just as the circadian clock has been identified in a range of organisms, so to has the temperature compensation mechanism. The identified mechanics, however, vary from organism to organism. The PER protein in Drosophila, for example, dimerises at a PAS domain. An additional domain of the PER protein has been found to also bind to this site. When the PS domain was mutated, the ratio of these interactions was altered and temperature compensation was lost (Huang et al. 1995). Similarly, temperature compensation in Neurospora depends on FRQ stability (Ruoff et al. 2005). This protein was previously discovered to exist as two isoforms (Garceau et al. 1997). Further research identified the balance of these two alternative splice variants was involved in temperature compensation (Diernfellner et al. 2007). This use of alternative splicing has successfully allowed temperature compensation to be modelled within the Neurospora circadian clock (Tseng et al. 2012).

1.1.4 – Molecular Mechanism

Circadian clock genes in the core oscillator of the clock can be differentiated by their phase, with a simple divide between genes that have a peak mRNA expression just after dawn (morning genes) and those with peak mRNA expression just after dusk (evening genes). These groups can then be further organized by exactly how long after dawn they reach maximum expression. Additionally, the effect of knocking out the gene and the importance of the gene under different environmental conditions can be used to place the gene within a network. This network attempts to inform experiments by suggesting interactions without the need for mathematical modeling or simulations.

1.1.4.1 – Circadian System

The clock can be split into three major sections. These relate to the core oscillator, the input into the oscillator that allows entrainment, and outputs

from the clock that show a strong circadian rhythm but have no mutant phenotype (Fig 1.1). Whilst this is a useful thought abstraction, it is possible for an input or output gene to also have a role within the core clock. For example, CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) is part of the central clock. However it also drives the expression of CHLOROPHYLL A-B BINDING 2/3 (CAB2). As such it could be classed as both a core oscillator component and an output component. Understanding the core oscillator in many organisms including plants has been the subject of substantial research efforts.. This part of the circadian system is believed to be the underlying timekeeper that causes the periodic cycling of other processes. Under constant environmental conditions, the clock continues to keep time with an approximately 24-hour output rhythm.

Despite the oscillator's ability to free run in constant conditions, tissues, cells, and even gene sets start to lose synchronicity, especially in the absence of environmental sucrose and light (Dalchau et al. 2011). To combat this, the input pathways of the circadian system act to co-ordinate the individual rhythms. A clear example of this is the effect of light/dark cycles on the circadian oscillator. Under constant conditions, traces of gene expressions appear to follow a sinusoidal pattern. However when they are exposed to light:dark cycles, many of the core oscillator components are seen to have rapid induction and/or degradation at the light:dark transition. This rapid induction allows the circadian clock to reset the clock based on external environment and maintains relative phase relationships of the different components.

The core oscillator also requires methods to transduce the oscillatory output to the rest of the plant. These output pathways allow for the rhythms to be communicated in a manor that can be further regulated without also affecting the core oscillator. These components also provide good candidate genes for monitoring the clock dynamics without having to directly interfere with the central oscillator.

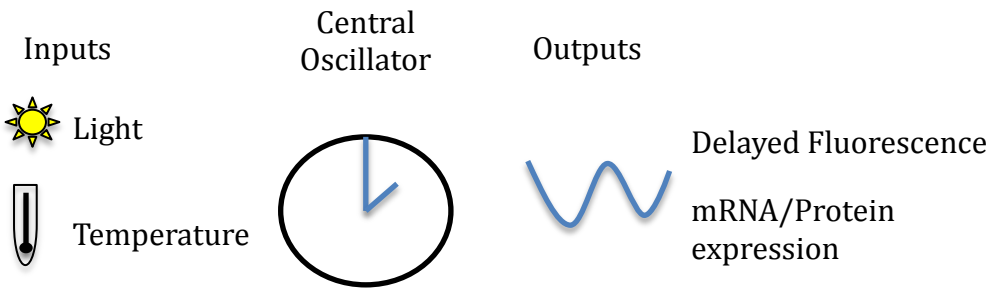


Figure 1.1 Basic components of the circadian system. Diagrammatic view of the three major sections of the circadian clock. Environment cues (such as light sensing) act as inputs, entraining the central oscillation, which then controls the expression of various output processes (such as leaf movement, delayed fluorescence and gene expression)

1.1.4.2 – Core Components

Using a wide range of biochemical and genetic approaches on a wide range of biological organisms, many individual genes have been identified as being important to different sections of the circadian system. For example, within the fungus *Neurospora crassa*, it has been discovered that the mRNA and protein forms of FREQUENCY (FRQ) are the primary elements of its clock, with the mRNA driving protein production and the protein inhibiting mRNA production (Feldman & Hoyle 1973). This is facilitated by 2 transcription factors – WHITE COLLAR 1 and 2 (WC-1, WC-2) – which promote the translation of FRQ (Crosthwaite et al. 1997)(Fig 1.2A). This simple feedback loop provided the basic model for circadian clock architecture. In the higher plant Arabidopsis, it was found that multiple genes were required to create a network that resembled the biological data (Alabadí et al. 2001). However, the overall architecture was closely conserved, with one set of genes activating another, which in turn represses the original set. The first clock gene identified in plants was TIMING OF CAB EXPRESSION (TOC1) (Millar et al. 1995). Several years later, two genes were identified as repressing *TOC1,* namely the single MYB-repeat transcription factors CCA1 (Wang et al. 1997) and LATE ELONGATED HYPOCOTYL (LHY) (Schaffer et al. 1998). These 3 genes form a loop similar to the one defined for *Neurospora* (Fig 1.2B (Alabadí et al. 2001)) and provided a first abstracted model of the plant clock. However, over time additional genes and loops were added to the circadian oscillator. These loops added other core circadian genes, including other members of the PSEUDO RESPONSE REGULATOR (PRR) family, of which TOC1 is a member (Makino et al. 2002), as well as GIGANTEA (GI) (Locke et al. 2006). Additional loops have been required to accommodate a growing body of mutant data. Many of the core clock components can be knocked out individually without causing complete arrhythmia, although clock controlled gene expression is generally severely effected (Alabadí et al. 2002; Somers et al. n.d.).

As additional components were added to the circadian network, the role of some components was only important at a protein level. With this increase in

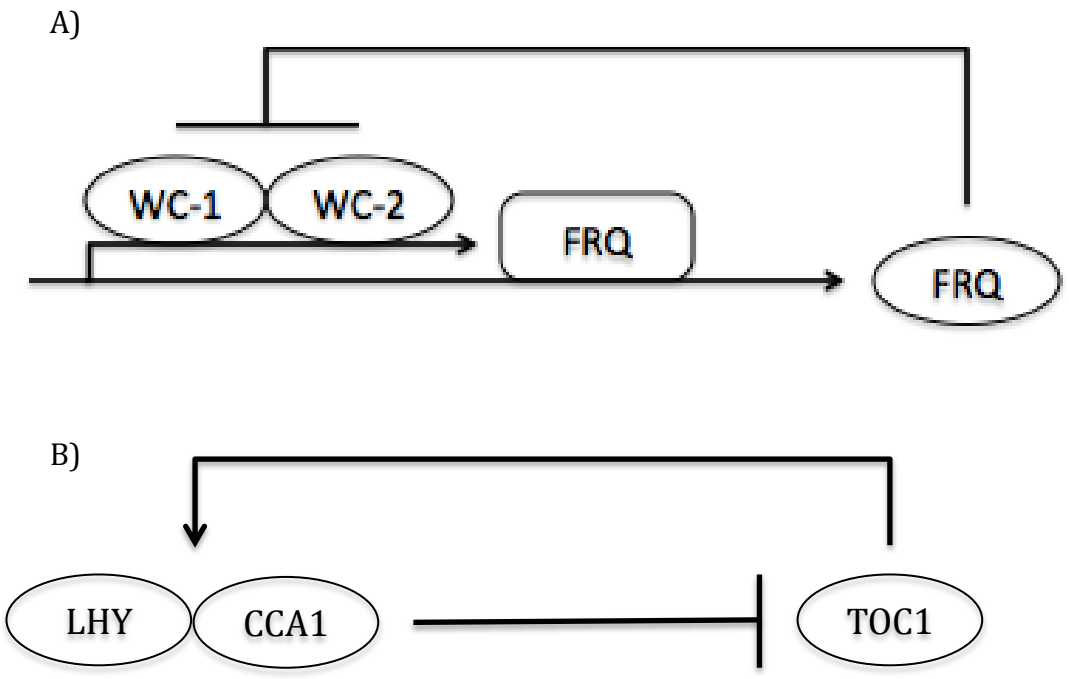


Fig 1.2 Simplified models of the circadian clock found in A) Neurospora crassa *and B)* Arabidopsis thaliana*. In* N. crassa*, WC-1 and WC-2 form a complex promoting FRQ transcription. FRQ protein then inhibits the action of the WC complex. In comparison, the* A. thaliana *clock consisted of 3 genes interacting with each other. TOC1 promotes the transcription of CCA1 and LHY. These proteins then inhibit the transcription of TOC1 protein.*

protein components of current circadian clock models, the need for mRNA for consistent rhythms came into question. Within cyanobacteria, a family of Kai genes are essential for clock operation (Ishiura et al. 1998). One of these proteins, KaiC, undergoes cycles of phosphorylation that are circadianly regulated. Recently, it was found that these cycles could persist without mRNA and in the presence of transcription or translation inhibiters (Tomita 2005). This protein-only circadian clock could also operate in constant dark (DD) and constant temperature conditions (Hosokawa et al. 2011), and to an extent have temperature compensation (Nakajima et al. 2005). The role of such post-translational mechanisms have been found to be more highly conserved than genes within the transcriptional clock, although the mechanism of protein modification can vary (O'Neill et al. 2011).

1.1.4.3 – Inputs and outputs of the Circadian Clock

The additional genes that have gradually been added to form a complex model of interlocking loops, including PRR9/7 and GI, are now also thought to function in the input of external cues into the circadian clock (Eriksson et al. 2003; Martin-Tryon et al. 2006). However, clock input pathways also require additional components. Within light entrainment, for example, there are two important families of genes. Phytochromes (PHYs) are crucial for regulating how plants sense red light (Reed et al. 1994). Similarly Cryptochromes (CRYs) play a vital role in sensing BLUE light (Yu et al. 2010).

In addition to core oscillator genes and genes involved in cueing the clock to external signals, there are many direct outputs of the clock. These outputs can be measured to understand what is happening to the clock at a systems level rather than at the level of single components. As the clock has been found to contain more complex interconnected and redundant loops, the use of clock outputs have become of greater importance. Two genes that have been used extensively are CAB2 and COLD CIRCADIAN RHYTHM AND RNA BINDING 2 (CCR2). CAB2 is predominately promoted by genes in the evening loop whilst

morning loop genes regulate CCR2 (Millar et al. 1995; Harmer 2000). Thus, by considering these two reporter genes, an overall understanding of how major sections of the core clock are changing can be achieved. However, this still requires either the transformation of plants with a promoter:reporter construct, or a destructive assay.

1.1.5 – Measuring the Circadian Clock

One important and powerful method for investigating components of the circadian clock involves using methods that monitor how clock-controlled gene expression changes over the day. These methods are focused on capturing data frequently without causing damage or stimulus to the plant. This allows the same plant to be used at each time point, providing an accurate measure of the gene expression profile.

A key method for gathering this type of data involves using the luciferase reporter gene as a visible measure of gene expression (Millar et al. 1992). This method works by fusing a gene's promoter region to the firefly luciferase (LUC) gene and introducing this transgene into the genome of the host plant. When such transgenic lines are exposed to the luciferase substrate, the LUC protein emits light. The intensity of this emitted light is directly proportional to the amount of LUC protein present, allowing a quantifiable measure of the promoter's activity. Thus, a non-destructive measure of promoter activity can be made via the quantity of light emitted at different times. This technique has been widely used in circadian biology as it allows real time measurement of gene promoter activity in a high throughput manner (Millar et al. 1995). The result of this screen is representative of the promoter activity although time is slightly skewed due to luciferase protein dynamics. However, this effect is the same for all genes, thus in any comparison between genes this effect can be ignored. Additionally, this effect is temperature sensitive, as such a back calculation to the promoter activity needs to be made to accurately compare expression profiles across temperature conditions (Costa et al. 2014).

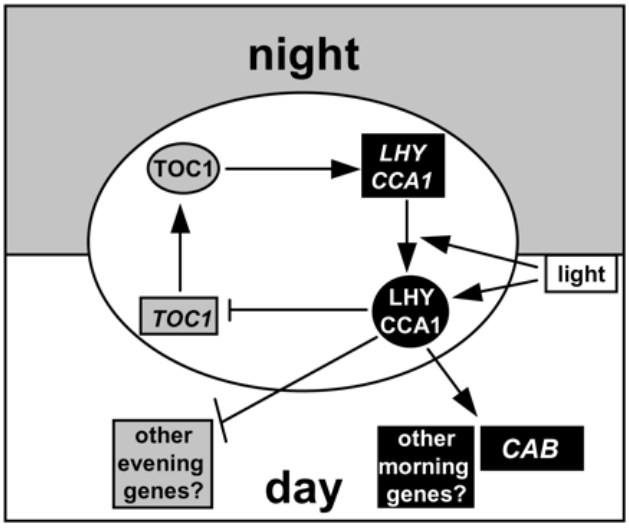
Although luciferase screens are very fast and high throughput, the generation of transgenic promoter::LUC plants is time consuming. Luciferase screens require the insertion of a marker and multiple generations to generate single loci insertions. This requires months of work and is impractical for screening large numbers of mutants for general circadian effects. Instead it is more practical to use a naturally occurring marker for circadian rhythms. Photosystem II provides such a marker via delayed fluorescence (DF). DF is generated due to photon emission from chlorophyll A caused by the over excitation of the photosynthetic electron transport chain (Rutherford et al. 1984). DF was found to be under circadian control, with the amount of light emitted relating to the phase of the clock (Gould et al. 2009). As such, measuring this photon emission at regular intervals provides a non-invasive assay for monitoring circadian rhythms in plants. This method has the advantage of being applicable to any mutant genotype without the need for transformation. However, it only serves as a general readout of circadian clock output, and does not report on the rhythmic regulation of specific core clock components.

Other assays to monitor the output of the central circadian clock in Arabidopsis are available. For example, leaf movement is a non-invasive assay that reports clock function (Edwards & Millar 2007). Over the course of a day, leaves move up and down in response to changing light stimuli. It has been shown that when core clock genes are knocked out, leaf movement rhythms are perturbed. The circadian control of gene transcript levels can be monitored microarrays and/or RNA-seq tagged proteins. Methods are also available to monitor protein abundance; for example immunohistochemistry (Matsuo et al. 2003). Furthermore, confocal microscopy can be used to monitor fluorescently tagged proteins. This provides information on protein quantity, and also includes information on the sub-cellular and cellular location as well as the cell type a protein is present in.

1.1.6 – Abstract Model of the Circadian Clock

Analysing data from the types of circadian biology screening methods described above provides valuable information about how gene expression profiles fit together. Using information about the phase of peak activity of each gene, it is possible to produce a representative network model of the system by arranging genes in the order that they have greatest active. This biochemical approach uses information gathered from multiple experiments to determine how genes interact. The first model of this type for Arabidopsis was produced by Alabadi et al., (2001) and consisted of just 3 genes (Fig 1.3A). As additional genes were identified as interacting with these components and having circadian phenotypes, they were placed into the network using biological information to determine how they connected to existing components. For example, given the acute induction of both CCA1 and LHY around dawn these are the first genes within the morning loop. PRR9 and then other members of the PRR family follow these two genes creating a gene cascade, identified by the time of peak expression. This is mirrored in the evening loop, where TOC1 peaks around dusk, along with GI and members of the evening complex (EARLY FLOWERING 3 (ELF3), ELF4 and LUX ARRHYTHMO (LUX)). As such these genes can be represented as two connected circles, where one is dominant during the day, and the other is dominant at night (Fig 1.3B). However, as these models expand to include larger datasets and more components, producing testable hypotheses and determining how well the model matches the biological system becomes increasingly difficult.

A)



B)

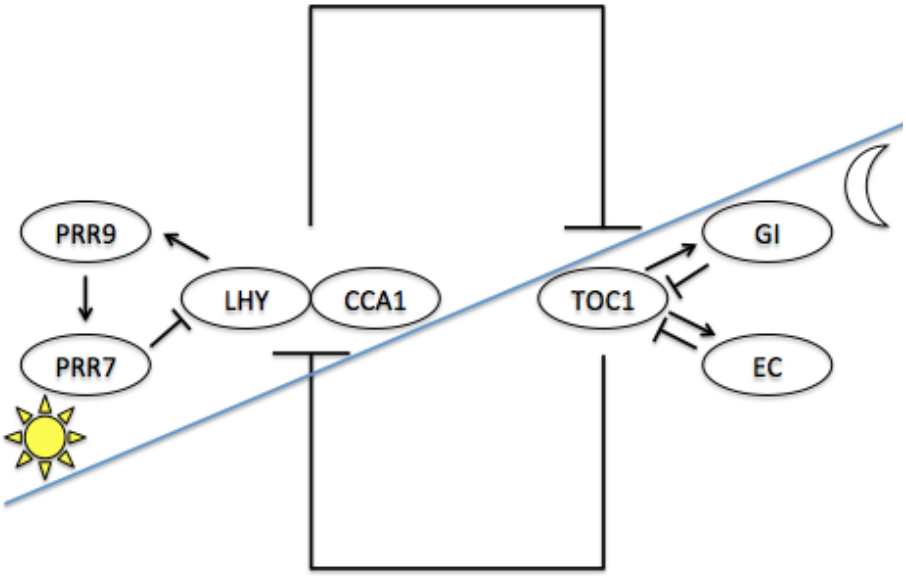


Figure 1.3 Abstract model of the circadian clock. A) Original model produced by Alabadi et al. 2001. Reprinted with permission from AAAS B) Expanded model to include additional components. LHY, CCA1 and PRR genes form a gene cascade around dawn. As the expression of the morning genes reduces, TOC1, GI and members of the evening complex start to peak, coinciding with dusk. These genes also inhibit the morning genes, preventing them from peaking until after the evening genes are under repression by the evening complex. The day/night divide it represented by a blue line.

1.2 – Systems Biology

Systems biology is the integration of experimental approaches and theoretical analysis to provide deeper insights into complex biological systems (Kitano 2002). This multi-disciplinary research approach uses mathematical and computational methods to analyse the biological data. With this added analytical power, biological experiments can be more holistic, generating information on a greater number of components. With this extra information and analysis, synthetic systems can be made more intricate in order to capture more of the biological dynamics (Snoep & Westerhoff 2005).

This approach is useful in a wide range of biological research (Noble 2006). It has been particularly important in furthering the understanding of systems that produced complex data sets that were difficult to interpret intuitively. It has been especially powerful in situations where the data produces a network that can then be simulated to validate the network (Locke et al. 2006). In these circumstances, simulated data can be compared to experimental data as a measure of how well a computationally reconstructed system can explain the in vivo biology.

1.2.1 – Network Analysis

Simulations to compare what is known about a biological process against real data require a generalised model of the system. The first step in producing a model is to create a network of gene interactions. At its simplest level, this is built from an understanding of the dynamics that exists between different elements of the system. Using biological knowledge of how genes behave when another gene is overexpressed or knocked out, a list of which genes affect other genes can be generated. Using this list of which genes affect which other genes, and the sign of the proposed interaction, a simple connected graph of provides sufficient data to create a basic model of the system. These models work by having nodes representing the genes and edges representing the interactions

between those genes. An example of this is the original circadian network proposed by Alabadi et al. (2001). A network was created that attempted to describe the dynamics occurring between the components of the system, but no equations were used to define these dynamics. In general, this method creates a heavily simplified model, as it does not account for protein dynamics. The complex processes of transcription, translation, protein modification and both transcript and protein degradation of each product get summarised into a single component. Although removing so much biological information results in networks becoming symbolic rather than absolute, the simplification allows for rapid modeling of multiple conditions and easier identification of key nodes changing between conditions (Akman et al. 2012).

One way of identifying how networks can be designed is by clustering genes based on their expression profiles. This provides multiple useful sets of information. Firstly, it helps to simplify large data sets by grouping genes whose expression over time are closely related, allowing multiple genes to be modelled as a single entity. Using a hierarchical cluster technique, potential cascade pathways can be identified based on the progression of maximum expression across the clusters. Different clustering techniques also allow the user to focus on specific elements of the dataset rather than having to look at the whole thing at once. These different elements can then be combined using consensus clustering, providing yet more ways to classify gene similarities and potential interactions (Breeze et al. 2011).

There are a number of available network inference packages that take time course data and use it to infer the underlying network that created the data (Penfold & Wild 2011). Software packages differ depending on the algorithm and technique they use to reproduce a network. Some attempt to identify which set of genes best explains a specific expression profile, and generate a list of gene interactions onto each node in turn. Others are concerned more in building a network as a whole, generating models at each step rather than considering genes individually. The algorithms used to determine whether one gene regulates another are also diverse. The algorithm can look for a simple linear

relationship, as in the case of VBSSM (Beal et al. 2005), or a more dynamical relationship as in the Gaussian process used by CSI (Penfold & Wild 2011). These network inference packages have primarily been designed for signal transduction pathways (Breeze et al. 2011; Windram et al. 2012), however it has been shown that these methods can be applied to an oscillating system as well (Penfold & Wild 2011).

1.2.2 – Modeling Networks

As circadian clock network models become more complex, the question of how specific genes are expressed over time becomes more complex and less intuitive to the classical biological way of thinking. A solution to this is to create a model of the network. Depending on the quantity of data available, and how complex the model needs to be, a range of different techniques can be used. Many models utilise ordinary differential equations (ODE's). These equations consider the rate of change of each element with respect to time. Models of this form are usually variations of Goodwin's oscillator (Goodwin 1966). The system that was derived from this oscillation takes the form:

$$\frac{\delta u_i}{\delta t} = f_i(u_{i-1}) - k_i u_i \qquad \text{Eqn. (1.1)}$$

where u_i is modulo n. This means that in a given system of n components, there are n equations (one for each component). These equations describe how the amount of component x changes over time in response to the levels of the other components in the network. For example, in the Locke et al. 2006 model, there are 16 equations with a form similar to Eqn. 1.1.

However, these equations frequently contain a large amount of variables. For each component there is a rate of generation (i.e. transcription and/or translation rate) as well as a rate of degradation. In addition, any interaction with another gene has an additional rate coefficient as well as a Michaelis-Menton constant (Michaelis & Menten 1913) to the power of a Hill coefficient

(Hill 1910). Because of this, within the 16 equations that described the Locke et al. (2006) model, there were 80 kinetic constants to be parametised. This number has increased as the circadian clock model has evolved to include more components (Pokhilko et al. 2012). Instead of optimising an increasing number of parameters, it is often simpler to consider the effect of one gene on another as a single variable.

A simple technique for developing this single variable type of modelling is Boolean modelling. Every gene is assigned a probability value in the interval [0,1]. At each time step, each gene is realised by comparing it's probability to a random number, also in the interval [0,1]. If the random number is equal or lower than the genes probability value, the gene's state is called on (or 1). Conversely if the random number is greater than a gene's probability value, it is determined to be off (or 0). Each gene's probability value is then updated using the variables between the gene and any other genes that have an impact on the genes activation that are currently thought to be active (on). This method allows many statistical analyses of network data to be quickly transformed into a model without the need of parameter estimation or optimisation.

1.3 – Modeling the Circadian Clock

Given the importance of the circadian clock in regulating the timing of many processes within a plant, using systems biology to understand the mechanisms by which it is regulated becomes crucial. Whilst abstract models describing how major components interact to occur anti-phase is useful, it lacks the specificity needed to understand how the measured data is being produced *in planta.* When a core component is altered in the plant, many effects occur because of it. Using an abstract model, it is often difficult to predict the effect on the other components, especially in a system like the clock where there are so many feedback loops. Using a mathematical model, however, the change to the physiology can be simulated and results hypothesized. This can help select which experiment is most likely to produce the desired results.

1.3.1 – Evolution of the Circadian Clock Model

The model of the plant circadian clock has gone through a number of iterations and developments in the last 8 years as more experimental data uncovered extra components and new connections. One of the earliest models was published in by(Locke et al. 2005). It consisted of two genes – LHY and TOC1 – and two hypothetical proteins – X and Y (Fig 1.4 A). Simulations of this model predicted a specific mRNA expression profile for gene Y. This included an acute peak at dawn as well as a more conventional peak around dusk. Using this information, an experiment was designed to look specifically for this expression profile by screening at a higher time resolution at dawn. Through this, GI was identified as a candidate for protein Y. This was then incorporated into the model, as well as an additional loop including the PRR's that was discovered at the same time (Fig 1.4 B). This increased model had six genes – CCA1, LHY, PRR7, PRR9, GI, TOC1 – and two hypothetical proteins, X and Y. Although GI captured a large proportion of the hypothetical protein Y, it was not sufficient to fully explain the mutant data. This simplistic 3-loop model was capable of capturing the majority of the published data, generated under standard growth

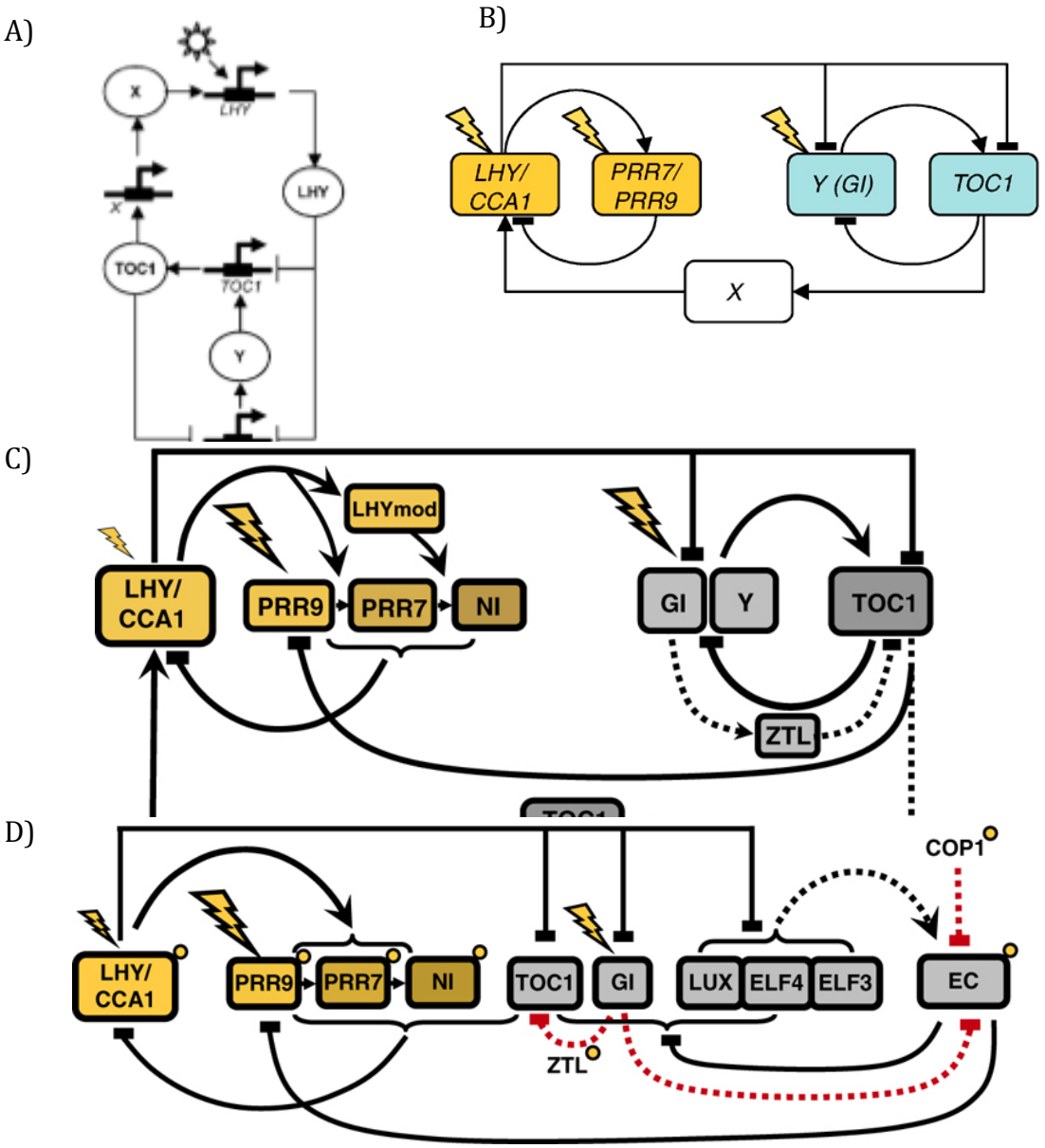


Figure 1.4 Evolution of the Arabidopsis Circadian Clock Model. Visual depictions of the various incarnations of the circadian clock model from A) a 2-loop model (adapted from Locke et al. 2005), B) a 3-loop model (adapted from Locke et al. 2006), C) an expanded 3-loop model (adapted from Pokhilko et al. 2010) and D) the repressilator (adapted from Pokhilko et al. 2012).

conditions. However, inclusion of hypothetical components suggested it was missing important information. Furthermore, light input to the clock was relatively simplistic in this model, with important nodes identified with lightning bolts. However, simulations using light regimes that were not 12 hour light:12 hour dark (12:12) did not match data as well as later models do.

This model was subsequently expanded by (Pokhilko et al. 2010) to better represent the morning loop, and to include ZTL within the evening loop. X was replaced with a modified form of TOC1 and a new unknown, NI, was included (Fig 1.4C). The split of PRR9 and 7 along with the addition of NI, of which PRR5 was thought to be a likely component, allowed the model to fit the data better, where there was a small but significant delay in the peak of these genes. This model was further developed by (Pokhilko et al. 2012) making use of new data on a multi protein complex referred to as the evening complex. This complex removed the need for component Y. Similarly the discovery that TOC1 binds to the promoters of LHY and CCA1 and represses their expression (Gendron et al. 2012) was incorporated into the model (Fig 1.4D). This model also included a more complex relationship between light and the circadian clock. Light interacts with the model at more points and in a more specific manner.

1.3.2 – Limitations of Models

Despite the improvements that have been made to the plant circadian clock models, they are still limited by the assumptions and data used to create them. Samples used to generate quantitative data are not only multicellular, but often contain several tissue types and even groups of plants. Thus, the model being produced is at best an average model. This leads to complications when fluorescent protein constructs under confocal imaging show that the clock in different tissues are often asynchronous. The problem with this is best shown by whether a plant is grown on sucrose or not. Plants grown in the dark without sucrose exhibit a rapidly dampening circadian rhythm when compared to plants on sucrose in the dark. However, single cell imaging studies show that the clock

does not dampen, but rather individual cells become asynchronous with neighbouring cells, making the pooled sample arrhythmic as a whole (personal communication, Gould and Hall).

Furthermore, whilst light input is modelled in the clock, there is no mechanism by which other stimuli such as temperature may be considered. As such, current models may be good for looking at a whole plant grown on sucrose at 22˚C, but become increasingly misleading under other conditions are changed. Adding Arrhenius functions to light inputs has been shown to produce a model that matches many of the gene period differences that are produced by growing plants at different ambient temperatures (Gould et al. 2013). However, the precise mechanism that allows themocycles to entrain the clock, as well as maintain temperature insensitivity has not yet been modeled effectively. The current best model also attempts to model CCA1 and LHY together as a single component. Considering these genes are known to perform partially redundant functions, this perhaps is not surprising. LHY has been shown to have a more critical role at high temperatures, whereas at low temperatures it was CCA1 that was dominant in controlling circadian rhythms (Gould et al. 2006).

1.3.3 – Simplifying the Clock for Modeling

In addition to the problems linked to the types of assumptions mentioned above, adapting the clock for various environmental conditions proves difficult due to the number of elements that need to be modelled. This then leads to a large number of parameters or connection strengths that need to be optimised or calculated. For many of the genes in the plant circadian network, we know values for the mRNA level, the cytoplasmic protein level and the nuclear protein level. These all have a basal rate of production and degradation as well the change in rates caused by interactions with other genes. Much of the research into expanding the plant clock model involves adding components rather than an option to simplify the number of interactions. However, previous models have successfully simulated data without the need for protein data. As such,

only using mRNA abundance data to create a model, whilst simplistic, should be capable of capturing the major aspects of the network and provide a simpler skeleton to fit additional genes in (Akman et al. 2012). This will have parts missing, such as the evening complex, but just as X and Y have previously been used to explain a missing component or connection, so too can a representative variable (i.e. Z) be included to represent a protein complex, or complexes, should it/they be required to construct a viable model. Alternatively, a variable time delay between different genes can be used to model any additional protein step detected for some genes in the network.

1.4 – Project Aims

This project aimed at investigating the mechanisms, pathways and genes that allow the plant circadian clock in *Arabidopsis* to compensate for changing ambient temperatures and maintain a constant rhythm. This was undertaken using microarray data and LUC data produced at 12, 17, 22 and 27˚C.

First, existing global gene expression data generated by using microarrays to assess gene transcript levels at four temperatures was analysed using the gene function of differentially expressed genes. This microarray dataset also included the *GIGANTEA*- (gi-11) mutant in the hope that a mutant for a gene known to have temperature compensation function might affects the response of the transcriptome to temperature. Genes that were found to have the greatest variability from the microarray analysis were screened for circadian phenotypes in null mutants (T-DNA tagged mutants of Arabidopsis) via using delayed fluorescence.

Secondly, promoter::luciferase reporter constructs were generated in a parallel project for 48 genes that had either known central oscillator circadian function, function in entraining the clock to external stimuli, or as a direct output from the clock. These promoter::LUC lines were screened at four temperatures under three light conditions in LD and LL conditions. This data was clustered independently for each unique set of conditions; different cluster results were compared to identify how genes were co-expressed and whether the topology of the network changed with temperature.

Lastly, a subset of the luciferase data was used to seed network inference software. Starting with the data generated at 22˚C, the inferred network was compared to existing models. Next, the data from the other temperature conditions was analysed to investigate how the network changed with temperature. To conclude this section, an attempt was made to model the network in order to develop an oscillating simulation.

Chapter 2 – Materials and Methods

2.1 – Computational Methods

Computational methods were performed using either R (v2.15.1) or Matlab (2011b/2012b) depending on where source code originated. When graphs were produced in R, the ggplot2 (Wickham 2009) package was used. Where possible, the latest versions of software packages available were used, unless otherwise stated.

2.1.1 – BioConductor

Microarray analysis was completed using the BioConductor v2.16.0 (Gentleman et al. 2004) plugin for R. The CEL files downloaded from NASC (Scholl et al. 2000) were loaded in using the ReadAffy() command, which reads all of the .CEL files within the working directory and creates an object containing all the information. This object was then normalised using GCRMA v2.28.0 (Wu et al. 2005) to normalise probe intensities between arrays as well as across each array. This normalisation also corrected the reported intensity based on the proportion of GC to AT residues of each probe. In addition, mas5calls (Gautier et al. 2004) was performed on the arrays to determine which genes were significantly detected in the experiment (present/absent call). This led to the removal of genes whose intensity was not significant compared to the added controls and likely was the result of background luminescence.

2.1.2 – BioDare

Pre-processed Luciferase (2.2.6) and Delayed Fluorescence (2.2.5) results were stored in a large database called BioDare (Moore et al. 2014). This online resource provides a central deposit for all of the ROBuST data. It also acts as an interface for multiple normalisation and analysis techniques. Throughout this project I used it to store time course data as well as detrending the data and normalising gene expression values.

Detrending the data removed the accumulative luminescence seen in time course data. This made the major component of the curves the oscillations rather than a luciferase effect. Similarly the normalisation technique used scaled each gene to oscillate between 0 and 1 to begin with, whilst maintaining dampening responses seen in later time steps. This allowed multiple genes to be easily compared to each other. Without amplitude being considered as a major difference.

2.1.3 – ReTrOS

Luciferase expression time series were passed through ReTrOS (Costa et al. 2014) to remove the effects of the LUC gene. As the reporter is measured by protein activity, the mRNA expression is slightly different to the light captured in the screen. Additionally, this delay in recorded expression from the actual mRNA level is temperature dependent. By using this software, expression profiles recovered under different temperatures can be more accurately compared. ReTrOS performed a back-calculation that subtracts this translation effect from the measured fluorescence. This subtraction is performed based on the temperature the experiment was conducted in, verified by experimental investigation.

2.1.4 – Cluster Methods

SplineCluster (Heard et al. 2006) was downloaded from http://www2.imperial.ac.uk/~naheard/splinecluster/ before being executed using CYGWIN. This process was automated using a simple R script provided by the University of Warwick that pre-processed the data, submitted the cluster command, and then performed the graphing script. This script was later adapted to also collate the results into a single file. The clustering command was performed using default parameters with the exception of the P-value, which

was set to 0.0001 after this was found to produce a more manageable number of clusters. FFT-spline (Liverani et al. 2009) was similarly completed using CYGWIN and a second R wrapper script. Again default settings were used as the submitted data set was similarly structured to the one previously used to optimise the algorithm.

2.1.5 – Gene Function Analysis

MAPMAN V3.5.1R2 (Thimm et al. 2004) was downloaded from it's website and analysis was performed using the automatically installed figures and the correct ATH1 array information. BiNGO v2.44 (Maere et al. 2005) analysis was performed on Cytoscape. This was done using the plant GO slim file and the entire Arabidopsis genome as a background. Differentially expressed gene lists were sequentially submitted for analysis. For both of these methods, a Benjamini Hochberg false discovery rate correction was applied within the software to reduce false positives.

2.1.6 – Network Inference and Modeling

Network inference was done using VBSSM and CSI software, both made available by the University of Warwick. VBSSM was run primarily using a GUI interface developed in Warwick although raw data was exported for analysis. This software was run initially by loading the normalized detrended data for the genes of interest into the code and using the default parameters. Investigation into the variation between seeds was performed by increasing the number of times the algorithm iterated. The number of hidden states, K, was determined by the software by testing for a K value between 1 and 20. The value with the optimal likelihood was selected.

CSI exists purely at a code level without any user interface. The EM algorithm was chosen for all experiments following the advice of its creator, Chris Penfold. By adapting the header code, CSI was able to accommodate the bigger data sets

being used in this investigation. This adaptation involved adjusting the variables informing the code of: the number of repeats, the number of genes, and the number of time points. Additionally, the code generating possible parental sets (PaSet) was adapted to run on the Liverpool Matlab license by changing the combntns() command with nchoosek(). These were essentially two identical commands that produced every possible unique combinations of x items from a list of y items. This code was also experimented with to investigate how many simultaneous connections should be considered.

Both software packages were run using Matlab 2011b or 2012b. Matlab 2012a did not support some of the functions, and there was no difference in the output of Matlab 2011b and 2012b. Because of this, calculation performed using the old version of Matlab did not have to be repeated on the newer version. Iterations of the network inference software were run on a server to rapidly perform the calculations in parallel, and reduce the overall run time. Modeling and simulations were created and performed in Matlab 2012b. Simulations were performed using a probabilistic Boolean algorithm (Savage et al. 2008) run in Matlab.

2.2 – Experimental Techniques

Experimental data was collected by ROBuST member Peter Gould at the University of Liverpool. Sue Bird and Dana MacGregor performed the construction of promoters fused with the Luciferase gene at the University of York, whilst Peter Gould and Jack Young performed the inclusion into agrobacterium and the dipping of plants at the University of Liverpool.

2.2.1 – Seed Stock

Mutants used in the delayed fluorescence screen (Chapter 3) were obtained through the NASC database. Collaborators at York University developed constructs used for luciferase screens (Chapters 4 and 5). Additionally, members of the ROBuST consortium working at the University of Edinburgh crossed luciferase markers into circadian mutants. These were created as heterozygous lines and were homogenised using multiple self-crossing performed by technicians at the University of Liverpool.

2.2.2 – Seed Sterilisation

All seed underwent gas sterilisation in a fume hood prior to cold treatment. This was achieved by placing the open Eppendorf tubes containing the loose seed into a large desiccation jar. Five hundred ml of reverse osmosis water was used to dissolve two chlorine tablets (CLO-TABS, Arrow Solutions). This was atomised using 5ml of hydrochloric acid and the jar sealed. This was then left for 3 hours before seeds were removed from the desiccation jar and moved to a sterile flow hood for an additional 30 minutes to remove remaining chlorine traces. 0.15% agar was then added to the tubes and seed suspended within it (modified from (Desfeux et al. 2000)).

2.2.3 – Plant Growth Conditions

Seed was moved to a constant dark 4˚C room for 3 days to promote stratification prior to planting. Three hundred μL of Murashige and Skoog media (MS) was pipetted into 96 well microtitre plates. Around 8-12 seeds were placed into each well before a second, empty microtitre plate was placed inverted on top. The two plates were then taped together with microporous paper tape and placed in a 22˚C room grown under 12:12 light:dark cycles of 80 $\mu mol/m^2/s^1$ in a Sanyo MLR350 plant growth chamber. These plates were left for 14 days for luciferase screens and 21 days for a delayed fluorescence screen.

2.2.4 – Circadian Screens

Circadian effects of genes were monitored using two different screens. When a LUC line existed, a luciferase screen was used to monitor the expression pattern of the gene (Southern et al. 2006). For mutants ordered from NASC, which had no inherent luciferase markers, a delayed fluorescence screen was used. Both screens were completed using the same apparatus (see Imaging System below). Similarly the raw output data was initially processed using the same methodology (see Image Analysis below).

2.2.4.1 – Imaging System

Circadian screens were carried out in Sanyo MIR-553 incubators (Sanyo Gallenkamp, UK) set to the desired experimental temperature. Image acquisition was done using a top mounted ORCA-II-BT 1024 16-bit camera (Hamamatsu Photonics, Japan) cooled to -80˚C (Southern et al., 2006). This setup allowed for the simultaneous imaging of six 96-well microtitre plates per cabinet. Lighting inside the cabinets (when needed) was supplied by red/blue light emitting diode (LED) arrays (MD Electronics, UK). These arrays were

controlled using WASABI imaging software (Hamamatsu Photonics, Japan) that also captured images from the cameras. Light intensity was set to a total of 40 μmol/m^2/s, either split 20/20 for red/blue light or all 40 μmol/m^2/s coming from a single light source, depending on the light conditions desired (Gould et al. 2006).

2.2.4.2 – Luciferase Screening

A single column of 8 wells in a microtitre plate was assigned to each transgenic plant being screened. This allowed biological repeats of up to 12 genes to be processed per microtitre plate. These plates were grown for 14 days in 12:12 LD conditions at 22˚C. 5 mM D-luciferin (dissolved in 0.01% Triton X-100) was applied using a fine spray on the 13th day in a sterile flow hood. Plates were moved to the imaging chamber at dawn on the 15th day after coming out of the cold room, and images were taken every 2 hours for the next 7 days. During the first two days the plants experienced standard 12:12 LD cycles, before then entering 5 days of constant light (LL). Images were taken by turning off the lights, waiting 5 minutes to remove plant auto fluorescence, and then capturing light emission for 20 minutes. This led to a plant experiencing 1 hour 35 minutes of light every 2 hours during light conditions. Cabinets were set to remain at a constant temperature of either 12, 17, 22 or 27˚C.

2.2.4.3 – Delayed Fluorescence Screening

Microtitre plates were set up in the same way as luciferase screens (2.2.4.2) but were left to grow for 21 days before being moved to imaging chambers. Additionally, each plate had a wild type control on it. Plants were left in constant red/blue light conditions from the start, with images being taken every hour. Image capture occurred by turning off the lights and immediately recording photon emission for 1 minute before lights were turned back on (Gould et al. 2009). This was done using cabinets set to maintain a temperature of 12, 17 and 27˚C.

2.2.4.4 – Image Analysis

The RBF images produced during the screens were reloaded into WASABI and converted into TIFF files. These were then imported into Metamorph 6.0 image analysis software (Molecular Devices). Regions relating to each well were highlighted, as well as four additional background regions. The colour intensity of these regions were then measured for each time and outputted to an excel file. These files were then uploaded to Biodare (Moore et al. 2013) where they were background corrected, detrended, and normalised.

Chapter 3 – Regulation of the Arabidopsis Transcriptome by Temperature

3.1 – Introduction

Recent research aimed at understanding how the circadian system is buffered to ambient temperature suggested that it is the circadian system as a whole that changed rather than a specific gene or mechanism (Gould et al. 2013). This system wide buffering may expand past the internal clock, and be a feature of a more widespread response to temperature. It is also possible that the plant wide response to temperature may be controlled via the circadian clock, given the circadian clock has already evolved a highly buffered mechanism to deal with ambient temperature change. To test this, a screen of a larger number of genes that better represents the entire transcriptome was performed at 12, 17, 22 and 27˚C. An efficient method to do this was to use microarrays on genetic samples collected at the different temperatures. Using the Affymetrix Arabidopsis ATH1 microarray provided by NASC, over 22,000 genes from the *Arabidopsis* genome could be screened at once.

In addition to wild type samples, microarray analysis was done on a GI knock out mutant, *gi-11* (Richardson et al. 1998). This mutant was discovered to cause a delay to flowering time when grown under long days, but has no effect under short days (Fowler et al. 1999). In addition, it has been discovered that this mutant is involved in the long-term sensing of sucrose, but not the short term (Dalchau et al. 2011). These studies also showed that GI is closely linked to the circadian clock. However, the *gi-11* mutant did not alter the expression of either TOC1 of CCA1 when grown at 17˚C in 12:12 LD cycles (Gould et al. 2006). When the same mutant is grown at either 12˚C or 27˚C, however, TOC1 and CCA1 amplitudes are both significantly reduced. This reduction in amplitude is also accompanied by a loss of temperature compensation, suggesting that GI is required to maintain circadian periodicity at extreme temperatures but not at 17˚C.

Using these ATH1 arrays, differential gene expression between 4 temperatures, 12, 17, 22 and 27˚C, and 2 genotypes, wild type and *gi-11*, were investigated. Plants were grown for 7 days in 12-hour light, 12-hour dark cycles (12L:12D) at 22˚C (80

$\mu mol/m^2/s^1$ from fluorescent strip lighting). They were then moved to the experimental temperature and constant light (LL) for 3 days. On the fourth day, plant material was collected every 4 hours for 24 hours creating 6 samples in each test condition. Equal concentrations of RNA from each time point was pooled to give one sample per test condition. Whilst this pooling strategy did remove any information on oscillatory rhythms, it also allowed fair comparison between gene expressions that may vary across different times of the day. This sampling was done in three independent experiments to act as biological replicates, and also with a *gi-11* null mutant, producing a total of 24 samples. RNA samples were used to synthesize labeled cRNA and hybridised to ATH1 arrays by the Nottingham Arabidopsis Stock Centre (NASC) microarray service.

From this investigation, it was discovered that the circadian clock is unlikely to control the plant wide response to temperature, despite its temperature compensation mechanism. Additionally, 68 genes were found to be differentially regulated with both temperature and in a *gi-* mutant. Of these, 13 were then shown to have temperature specific circadian phenotypes when they were knocked out.

3.2 – Pre-processing and Quality Checks

Raw .CEL files downloaded from the NASC database were imported into R using the BioConductor package (Gentleman et al. 2004). These underwent normalisation using GCRMA (Wu et al. 2005) as well as being tested for whether they were statistically present within the array using mas5calls (Gautier et al. 2004). Using the present/absent calls, genes were tested for their presence in each condition. By using the three replicates for each condition, it could be determined whether a gene was present in each experimental condition set. Only if a gene was called present in two of the three repeats was it classes as being present in that condition. In addition, genes that were not present in both genotypes were removed from further analysis. Genes did not, however, need to be classed as present at each of the temperatures. This resulted in 14679 genes out of the original 22749 being used in further analysis.

A principal component analysis (PCA, (Wold et al. 1987)) was then performed on the remaining data as a quality check (Fig 3.1). The first component showed that the major variation was between elements in the arrays. However, this variation was similar in all experiments. There was a subtle, but statistically significant difference between arrays performed on wild type plants and those performed on the *gi-11* mutant. There was no significant difference between the 3 repeats, or between the temperatures within this first component. The second component within the PCA acted to split the experiments by the temperature they had been performed at (Fig 3.1). The value of this component increased as temperature at which the plants had been grown decreased, with the largest difference being between the extreme temperatures (12˚C and 27˚C). The plants grown at each temperature were closely clustered on this component, with no systematic difference between the genotype or repeat number. This lack of systematic difference again supported the conclusion that there was no technical bias within the experiments. This analysis showed that the genotype of the plant and the growth temperature caused the differences seen in the transcriptome, and that there was no experimental bias between the experiments.

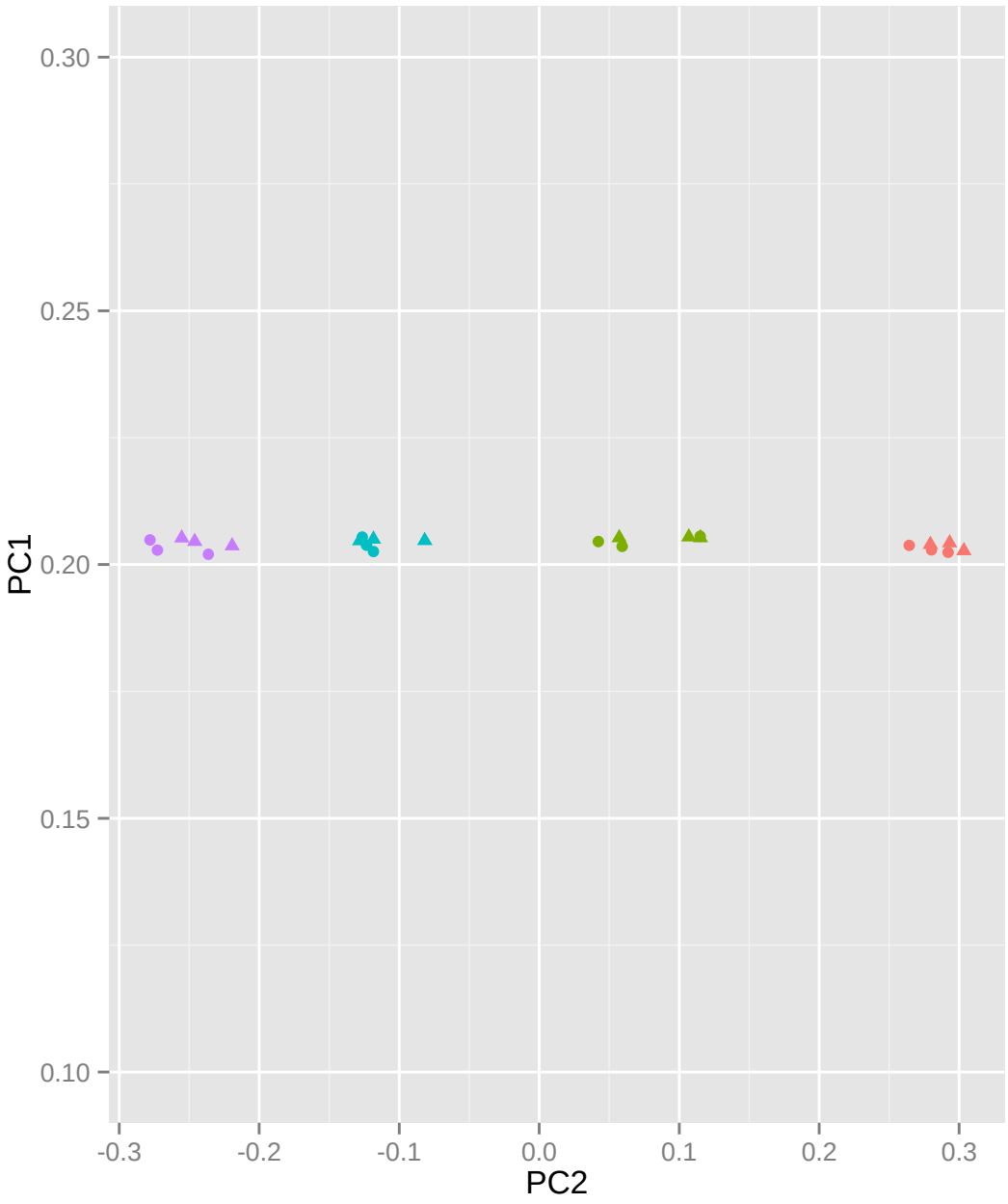


Figure 3.1 Principle component analysis of the 24 microarrays. These microarrays consisted of triplicate replication of four temperatures (colour) and 2 genotypes (shape). The first and second components were plotted for each microarray. Plants with a wild type phenotype were plotted with a triangle whilst plants with a gi-mutant phenotype were plotted with a circle. Colours of the points represent the temperature the plant was grown at: 12- red 17- green 22- blue 27- purple.

3.3 – GO Analysis of Differentially Expressed Genes in Wild Type Plants

Gene Ontology (GO) analysis is a method of identifying the putative or known functions of genes within an organism (Botstein et al. 2000; Ma et al. 2004). It consists of a tiered system of annotations within these three categories (Biological Process, Cellular Compartment and Molecular Function). This allowed the help investigation of te function of differentially regulatedgenes from the array analysis. Additionally, lists of genes could be used to identify GO terms that were being disproportionately identified in the list. Using this analysis on the set of genes being differentially regulated across the temperature range identified what processes were being manipulated as well as which specific genes.

For GO analysis, lists of genes with differential expression needed to be generated. Using a crude fold change between two temperatures might have removed genes with interesting expression at the other two temperatures. This could be countered by performing the analysis on every pairwise comparison. However, integrating the six results to understand what happened across the entire temperature range was susceptible to bias. Instead, clustering the gene set across the four time points identified genes with interesting responses to temperature even if they were just under a t-test statistic. This method was also less sensitive to false positives caused by a single point having an erroneous reading.

3.3.1 – Clustering Microarray Elements

There were several ways to cluster this multi-dimensional data. Methods such as k-means (Hartigan & Wong 1979) treated each data point as a separate entity, and formed multi-dimension ‘cages’ to split the samples. Others, like spline clustering (Heard et al. 2006), treated each data point as a point in a line, and attempted to cluster samples based on different elements within these

lines. Since this data was formed at four different temperatures equally spaced, using the latter method was likely to recover more genes with important expression patterns. One method of clustering sequential data for patterns in the curve between points was SplineCluster. SplineCluster fits a series of simple splines to a complex curve. These splines attempt to capture the most prominent characteristics of the curve. These characteristics can then be clustered together, resulting in a set of clusters based on the major dynamics of the curves.

Whilst the absolute values of the data have less impact on a spline cluster algorithm compared to a k-means algorithm, the difference in the raw value was likely to cause genes to be split. Thus genes that responded to temperature in the same way would have been split because of their raw value. However, this analysis was more interested in identifying genes whose expression changed over the temperature range rather than genes that have a higher absolute expression compared to other genes. As such, gene expression at each temperature was normalised to its expression at 17˚C. This temperature was chosen due to the effects seen in the circadian clock in the *gi-11* mutant (Gould et al. 2006). It was shown that at 17˚C, the mutant had no significant change in phase or period evident in circadian assays, although the amplitude was altered. In contrast, at the other three temperatures, there was a significant shortening to the period of the oscillations. However, due to the nature of the normalisation, using any of the temperatures as a control gives the same results. Normalised data was then passed into the SplineCluster software.

Within the software, a default P-value of 0.0001 was used to inform cluster thresholds. This resulted in a total of 39 clusters in wild type (Fig 3.2). Within these clusters, the first cluster contained all of the genes that on average showed less than 5% change across the temperature range. This equated to 10063 genes, over 68% of the genes used for cluster analysis. A further 18 clusters (coloured green in Figure 3.2) showed some change in expression, although not enough to be a significant trend. The remaining 20 clusters, coloured red and blue depending on if they increased or decreased in

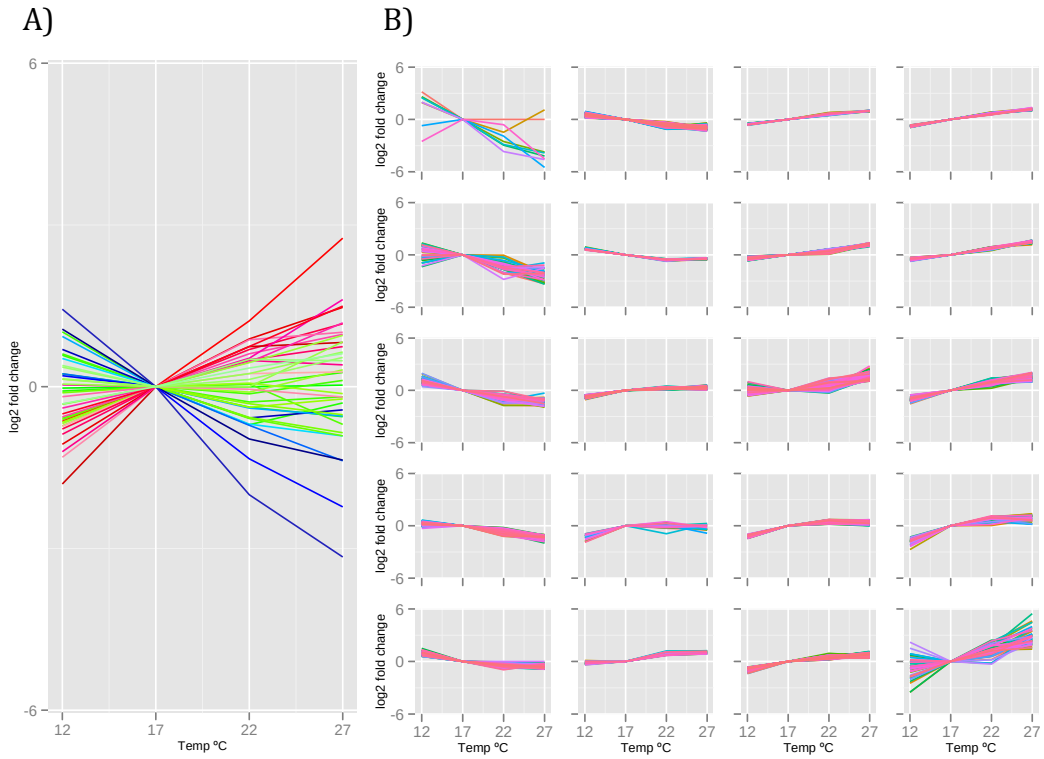


Figure 3.2 Log2 fold change in expression of clusters and genes at 12, 17, 22 and 27˚C compared to 17˚C in WT plants. (A) Plot of the average expression change with changing temperature in each of the 36 clusters. Clusters with significantly increased expression were coloured in shades of red. Clusters with significantly decreased expression were coloured in shades of blue. Remaining clusters had no significant change across the temperature series and were coloured in shades of green. (B) Each cluster that showed a significant fold change in response to temperature changes was then replotted on its own graph. Each graph represents a single cluster and lines within these graphs are individual genes. These graphs were ordered from the most significant reduction in expression to the most significant increase in expression.

expression as temperature increased, all showed greater than 2-fold expression change within the temperature range tested. These are shown in Fig 3.2 B, where each graph represents a separate cluster, and each line represents a separate gene. Individual clusters ranged in degree of variation, from greatly reducing with increasing temperature at the top left across to greatly increasing with temperature in the bottom right. These 20 clusters contained 1242 genes, 8% of the total gene set analysed. These are mostly located in the middle of Fig 3.2 B, where change across the temperature range was minimal. Only a small subset belonged to the outlying clusters where expression levels undergo big changes. Within the 1242 genes, 784 genes showed a clear increase with temperature (Supplemental Table 3.1) and 458 genes displayed a significant reduction in expression with temperature (Supplemental Table 3.2). Genes within these tables were ordered from clusters with the maximum fold change across the temperature range to those with the smallest change. There were no clusters where expression levels at the extremes were close but significantly different at mid-temperatures. This potentially suggests that genes were regulated in a linear manner across this temperature range.

From this analysis, 1242 genes that significantly changed in expression across the temperature range were identified in wild type plants, spaced across 20 clusters. Gene distribution across these clusters was not equal, however, with multiple clusters having very few genes. The clusters with only a few genes contained the genes with greatest change as temperature changed. Due to the nature of GO analysis, tests performed on gene lists with very few members often produced biased results. To avoid this, the 1242 genes were split instead into two lists, one for genes whose expression increased as temperature increased and one for genes whose expression decreased.

3.3.2 – BiNGO Results

To perform the GO analysis on these lists, BiNGO (Maere et al. 2005) was selected. BiNGO is a plugin for Cytoscape, a graphical display package designed

to look at interconnected variables (Shannon 2003), that displays the GO terms that appear in a list of genes more frequently than expected. The lists generated from cluster analysis were tested for global analysis. This was done using the GoSlim Plants annotation (Camon 2004; Miyama & Tada n.d.). GO slim is used because many of the more specific GO terms only appear once or twice in an organism. By using a GO classification with more general terms, issues with false positives caused by these underrepresented terms were avoided. Significant terms were drawn as a network to show how they fitted together in the hierarchical structure of GO. This could be done for every term in the GO slim list, or it could produce a figure that only displayed the significantly represented terms. BiNGO calculated whether a term was significant by comparing the frequency that a term appeared in the supplied list of genes to the frequency the term occurred in the supplied annotation database. This check for significance prevented terms that were highly prevalent within the entire organism as being falsely identified as significantly represented in the list of genes. In addition, a Benjamini Hochberg correction for false discovery was also applied with a P-value of 0.05 (Benjamini & Hochberg 1995).

This analysis was performed independently for each of the two lists of genes created from the results of clustering (Figure 3.3). This identified several specific terms with significant over-representation; most significant of these was ‘response to stress and stimuli’. Although identifying this term would be expected considering the temperature perturbations applied to the different plants, this result could be considered as a validation of the analysis. Many of the terms being significantly over-represented were found to be significant in both lists. With these processes being differentially regulated at both temperatures, there was a suggestion that there were sub groups within these families of genes that worked antagonistically to each other. This antagonistic relationship may indicate gradual change from a cold response to a heat response. For example, response to stress was found to be significant at both extremes. This might suggest that one subset of genes had increased expression at high temperatures whilst another subset had increased expression at low temperatures, rather than a simple response at one extreme. Also, the

Figure 3.3 GO slim over representation within differentially expressed genes. BiNGO visualisation of terms overrepresented in genes that were called up regulated (A) or down regulated (B) between plants grown at 12°C compared to those grown at 27°C. Significance was determined by considering the full Arabidopsis genome as a background and corrected using Benjamini and Hochberg false discovery rate. Shown were all terms significantly represented, as well as precursor terms in all three GO trees. These were coloured according to their P-value on the scale shown in the bottom right corner. White terms are those that were not significantly represented but were precursor terms for ones that are.

identification of the cell wall term in both lists was potentially interesting. Many genes located within the wall have been associated with cell growth (Cosgrove 2005). This suggested plant growth was being regulated by temperature, but not in a simple relationship whereby higher temperature equals more growth through up-regulation of growth genes.

From this BiNGO analysis, it could be seen that many of the genes that were overrepresented in the list of up-regulated genes (Fig 3.3 A) were also overrepresented in the list of down-regulated genes (Fig 3.3 B). However, there are some significant terms that were only defined as overrepresented in the list of down regulated genes. These were 'plastid', 'nucleolus', and 'cellular amino acid and derivative metabolic process' terms. These were all terms that were most commonly associated with either transcription or translation. This may suggest that, as temperature increase, genes that function in the transcription and translation pathways are being regulated in such a way to combat the increase in gene turnover caused by increased temperature.

Through GO analysis, several high order terms were identified which were potentially acting antagonistically to each other to provide a balance as temperature changed. However when the same GO term was found to be significant both with increased expression and decreased expression, it became harder to understand how the genes within this family were interacting, such as response to stress highlighted above. To further investigate this, the varying gene functions needed to be mapped onto metabolic processes to better understand how these processes were changing with temperature.

3.4 – MAPMAN Analysis of Microarrays

MAPMAN (Thimm et al. 2004) is a community driven tool that investigates how differential gene expression affects a process or network. By supplying a list of genes and their log2 expression fold change between two conditions, many biological processes can be explored to identify significant effects on pathways. Due to the statistics used in the software, providing information on a greater proportion of an organism's transcriptome produces more reliable results. In this analysis, MAPMAN was used to not only consider broad, general terms, but also investigate specific pathways, some of which contained detailed single gene connections. This ability to investigate what happens within a significantly affected process in higher detail meant it was possible to understand in greater detail some of the processes that GO analysis had previously identified as both increasing and decreasing with temperature.

3.4.1 – Global Overview

For MAPMAN analysis, the log2 fold change between the extreme temperatures (12˚C and 27˚C) was used. Cluster analysis showed there were few cases where using a mid temperature would produce a significantly different result, so only this one ratio was considered. This list was loaded into MAPMAN and analysis was performed using the top-level categories and a Benjamini Hochberg false discovery rate correction (Table 3.1). Each row represents a major category (referred to as a bin within the software) of plant biology. The elements column provided information on how many of the submitted genes were classified into each bin. The p-value for each bin describes how significantly that bin was differentially expressed in the data provided. From these initial results, several of the bins were found to have a significant p-value, e.g. bin 1 (photosynthesis). However, these top-level bins were not much more informative than the GO analysis was. To analyse what was happening in more detail, the sub bins that made up the bins shown were explored. However, many of the significant bins (such as photosynthesis) did not contain sub bins which were significant. This

Table 3.1. Differentially expressed functional classification between plants grown at 12˚C and 27˚C calculated using MapMan software. Significant bins are formatted in bold.

bin	name	elements	p-value
1	**PS**	185	**6.97E-03**
2	**major CHO metabolism**	81	**9.62E-04**
3	minor CHO metabolism	96	4.62E-01
4	**glycolysis**	56	**2.64E-03**
5	fermentation	13	5.84E-01
6	gluconeogenese/ glyoxylate cycle	9	3.82E-02
7	OPP	28	3.34E-01
8	TCA / org. transformation	63	2.60E-02
9	mitochondrial electron transport / ATP synthesis	108	1.50E-01
10	cell wall	299	9.58E-01
11	lipid metabolism	304	9.90E-02
12	N-metabolism	24	5.72E-01
13	amino acid metabolism	211	1.27E-01
14	S-assimilation	12	2.67E-01
15	metal handling	50	9.76E-01
16	secondary metabolism	280	2.12E-01
17	**hormone metabolism**	309	**1.78E-09**
18	Co-factor and vitamin metabolism	70	2.27E-01
19	**tetrapyrrole synthesis**	40	**2.52E-04**
20	**stress**	511	**3.21E-05**
21	redox	169	9.99E-01
22	polyamine metabolism	13	5.11E-01
23	**nucleotide metabolism**	131	**1.04E-03**
24	Biodegradation of Xenobiotics	19	1.50E-01
25	C1-metabolism	32	5.64E-01
26	misc.	835	1.01E-02
27	RNA	1719	5.60E-02
28	**DNA**	358	**3.92E-04**
29	**protein**	2333	**8.93E-08**
30	**signalling**	804	**1.37E-06**
31	cell	561	5.00E-01
32	micro RNA, natural antisense etc.	2	5.08E-01
33	development	466	9.06E-01
34	transport	695	1.25E-01
35	not assigned	4728	6.46E-01

meant that whilst the general term may have been significantly affected by the condition change, it was a general effect rather than a specific pathway or mechanism being differentially regulated.

3.4.2 – Pathway Specific Analysis

The main exceptions to this lack of low-level significance were the transcription, translation and degradation pathways (Fig 3.4, 3.5, 3.6). This correlated with the results attained from BiNGO analysis, as well as giving further information about the specific mechanism by which the gene expression profiles caused these processes to become differentially regulated across the temperature range. It can now be seen that specific transcription factor families were being up regulated in warmer temperatures (Fig 3.4). Fig 3.5, on the other hand, showed there were slightly more genes with increased expression at low temperatures in the transcription process. It also showed a clear increase in the plastidic translation at low temperatures. In addition, Fig 3.6 showed that the ubiquitin-binding genes were more active at low temperatures, but this was balanced by greater expression of other elements at higher temperatures. These results appeared to show that transcription, translation and protein degradation played a crucial role in maintaining biological functions across the changing ambient temperatures.

Figure 3.4 Differential regulation between 12°C and 27°C of transcription factors. Each clock of squares represents a family of transcription factors. Single squares represent single genes. If a square is coloured blue, then it was significantly more expressed at 12˚C, if a square is coloured red then it was significantly more expressed at 27˚C.

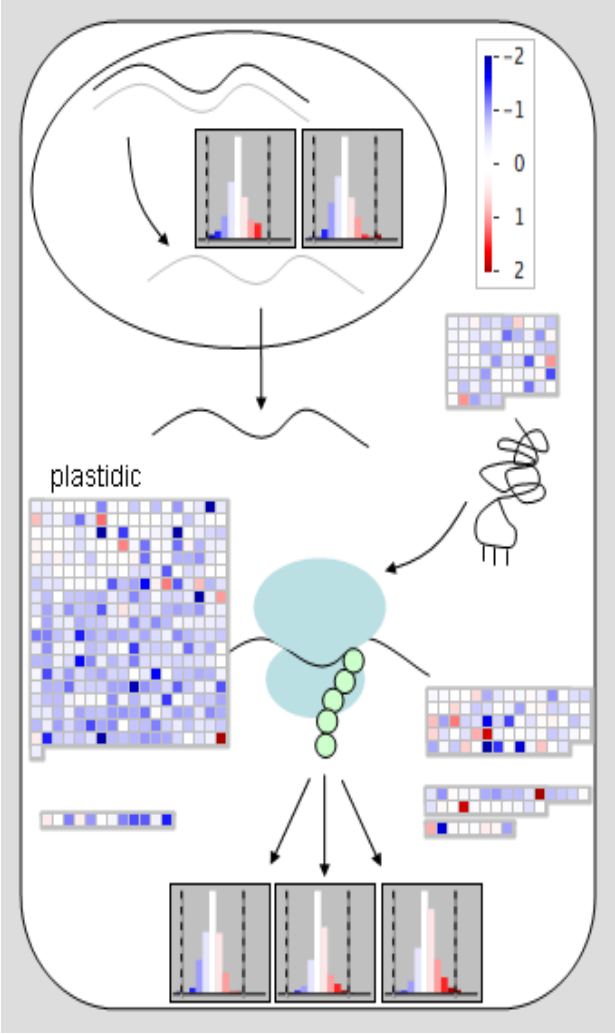


Figure 3.5 Differential regulation between 12°C and 27°C of the transcription and translation pathways. Histograms show the number of genes that were down regulated (blue) or up regulated (red) at higher temperatures for a specific function. Blocks also represent a specific category of genes, and individual squares indicate single gene IDs. If a square is coloured blue, then it was significantly more expressed at 12˚C, if a square is coloured red then it was significantly more expressed at 27˚C. The top oval represents mRNA translation, the block on the top left represents nuclear export, the bottom left and right blocks represent the translation pathway, and the bottom row of histograms represents post translational modification.

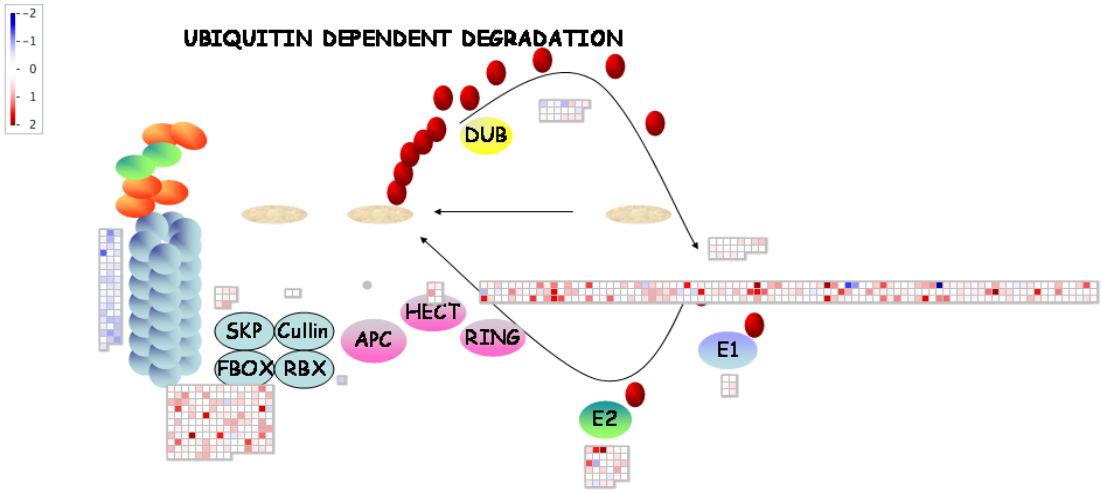


Figure 3.6 Differential regulation between 12°C and 27°C of the ubiquitin dependent protein degradation pathway. Blocks represent a specific category of genes, and individual squares indicate single gene IDs. If a square is coloured blue, then it was significantly more expressed at 12˚C, if a square is coloured red then it was significantly more expressed at 27˚C. Each block represents a major protein complex involved in this process.

3.5 – Analysis of the *gi-11* Mutant

Up to now, all analysis had been performed on the wild type microarrays. However this was only half of the data. The next step was to use the *gi-11* data set to see whether the mutation in this circadian temperature compensation gene caused any difference in the results. Circadian genes LHY and CCA1 had previously shown no differential expression is response to temperature change (Supplemental Table 3.6). However in the *gi-11* mutant, both of these genes were found to have a temperature response. In addition, gene expression was reduced at all temperatures for these genes, but most significantly at 12˚C and 27˚C. Thus, despite pooling data across a time series, this screen was able to identify how the circadian network changed with temperature with a reduced temperature compensation mechanism.

However, analysis using GO and MAPMAN on this second data set did not produce significantly different results. This showed that even when the temperature compensation was compromised, the same processes were being affected in a temperature dependent manner. This suggested that the majority of general temperature buffering effects were regulated by other factors than GI and that the circadian clock does not provide a generic temperature buffering system for the organism.

3.5.1 – Clustering Microarray Elements of the *gi-11* Mutant

Whilst MAPMAN and GO analysis did not produce any different results, this might have been caused by a more subtle effect in the mutant. To test this, the microarrays were clustered in the same way as the wild type ones, as described in Chapter 3.3.1 (Fig 3.7). Spline clustering produced 43 clusters, 27 of which were significantly differentially expressed. Although clustering produced more significant clusters, a similar number of genes being up and down regulated were found compared to the WT plant (767 and 496 genes respectively). In both these lists around 75% of the genes were also present in the equivalent WT list.

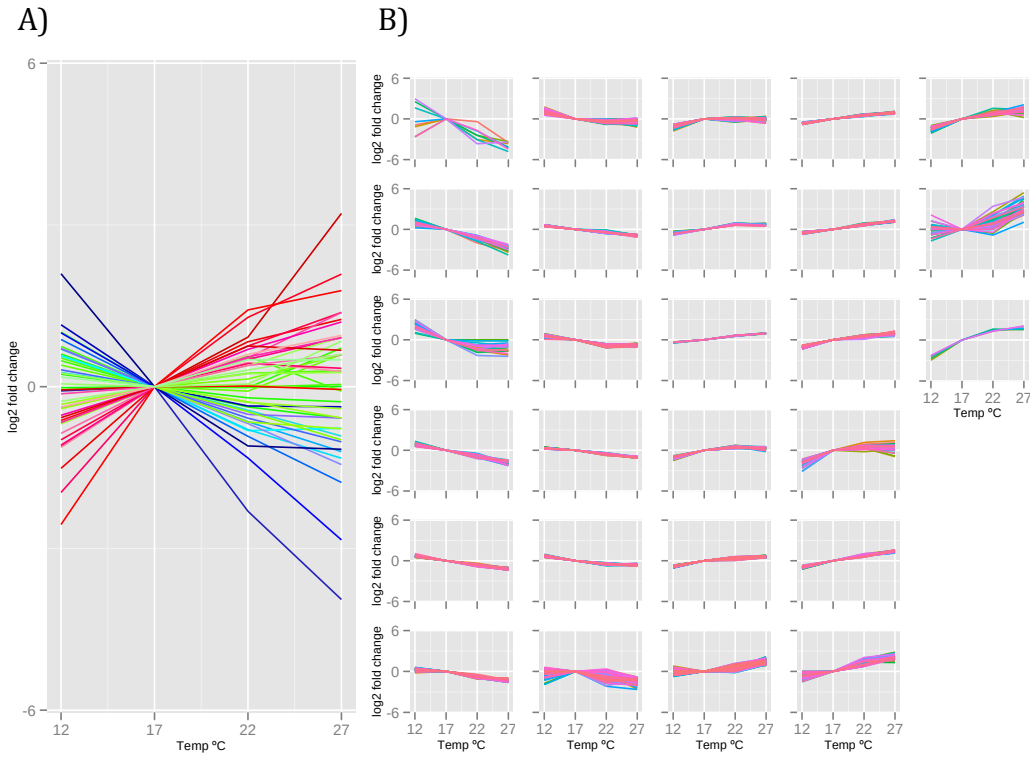


Figure 3.7 Log2 fold change in expression of genes at 12, 17, 22 and 27˚C compared to 17˚C in gi-11 mutant plants. (A) Plot of the average expression change with changing temperature in each of the 43 clusters. Clusters with significant increased expression were coloured in shades of red. Clusters with significantly decreased expression were coloured in shades of blue. Remaining clusters had no significant change across the temperature series and were coloured in shades of green. Each cluster which showed a significant fold change in response to temperature changes was then replotted on it's own graph (B). These were ordered from the most significant reduction in expression to the most significant increase in expression.

Considering the genes that were up regulated with temperature in both genotypes, 12 genes were found that showed at least a 50% increase in their temperature response in the mutant compared to wild type. However, no genes were found to have a 50% decrease or greater. Similarly, there were 14 genes that showed a 50% or greater increase in their temperature response out of the genes that decreased in expression with temperature, but none with a 50% reduction. This suggested that GI helped to control how some genes responded to temperature and limited their expression change, but this was only a small subset.

3.5.2 – Differential Expression Between Genotypes

To further examine how the mutant was affecting temperature compensation, genes were examined to determine which were differentially expressed between the genotypes at each temperature. A gene was declared differentially expressed if it showed a 1.5 fold difference between WT and the *gi-11* mutant (Supplemental Table 3.3). By comparing these lists, genes that were being differentially expressed only at extreme temperatures could be identified (Fig 3.8). Genes that were consistently differentially expressed were more likely to be involved in temperature independent processes, and not GI's role in buffering the clock. Through this, it could be seen how many genes were being up and down regulated at each temperature in the mutant. Interestingly, most of the genes identified as being differentially expressed in the *gi*-11 mutant occurred only at one of the extreme temperatures. There were also marginally more genes that were down regulated in the mutant than were up regulated. In addition, twice as many genes were differentially expressed at either 12˚C or 27˚C compared to genes differentially expressed at 17˚C. GO analysis for the genes differentially regulated in the *gi-11* mutant at 12 and 27˚C found that the 'response to stress and stimuli' term was significantly overrepresented in both genes classed as up regulated and those classed as down regulated. In addition, the transcription factor activity term was significantly over represented in the list of genes down regulated in the *gi-11* mutant at both 12˚C and 27˚C. The

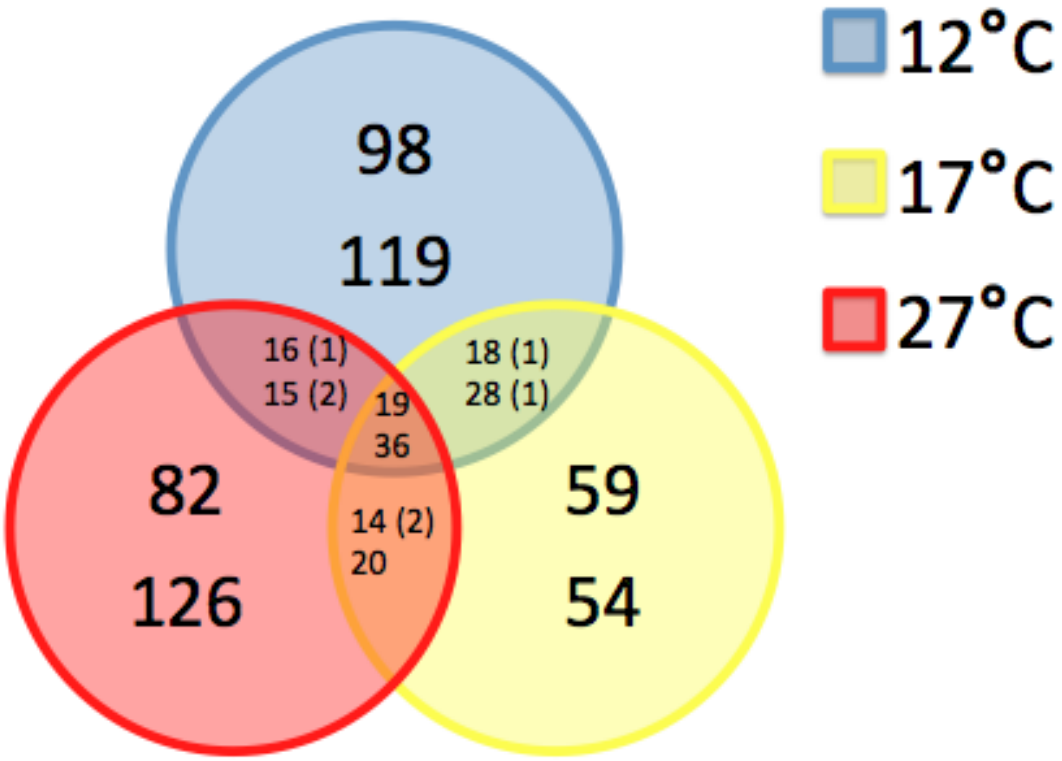


Figure 3.8 Effect of GI knockout on gene expression. Venn showing the number of genes differentially expressed in a gi-11 mutant compared to WT at each temperature. The top number of each section was the number of genes up regulated in the mutant; the bottom number was the sum of genes down regulated in the mutant. Numbers in brackets were where genes have the opposite regulation in the higher temperature compared to the lower temperature. For example, in the 17°C and 27°C section the bracketed 2 refers to two genes that are up regulated at 17°C, but down regulated at 27°C.

transcription term was also significantly over represented in genes down regulated at 12˚C in the mutant. Genes annotated as occurring within the plastid were over represented within the list of genes that increased in expression in the mutant at 12˚C. These findings closely matched the results of GO analysis on the wild type microarrays. This suggested that GI, and perhaps temperature compensation, acts to mediate the plants natural responses to temperature and limit the effect.

3.6 – Delayed Fluorescence Screen

After observing that similar results were found when analysing effects of changing temperature and knocking out a core temperature compensation component, the data was investigated to identify which genes were most sensitive to these conditions. These selected genes were then used to identify knock out mutants within the SALK database (Alonso 2003). Using these knock out lines, delayed fluorescence screens (Gould et al. 2009) were used to identify circadian rhythm phenotypes across the temperature range.

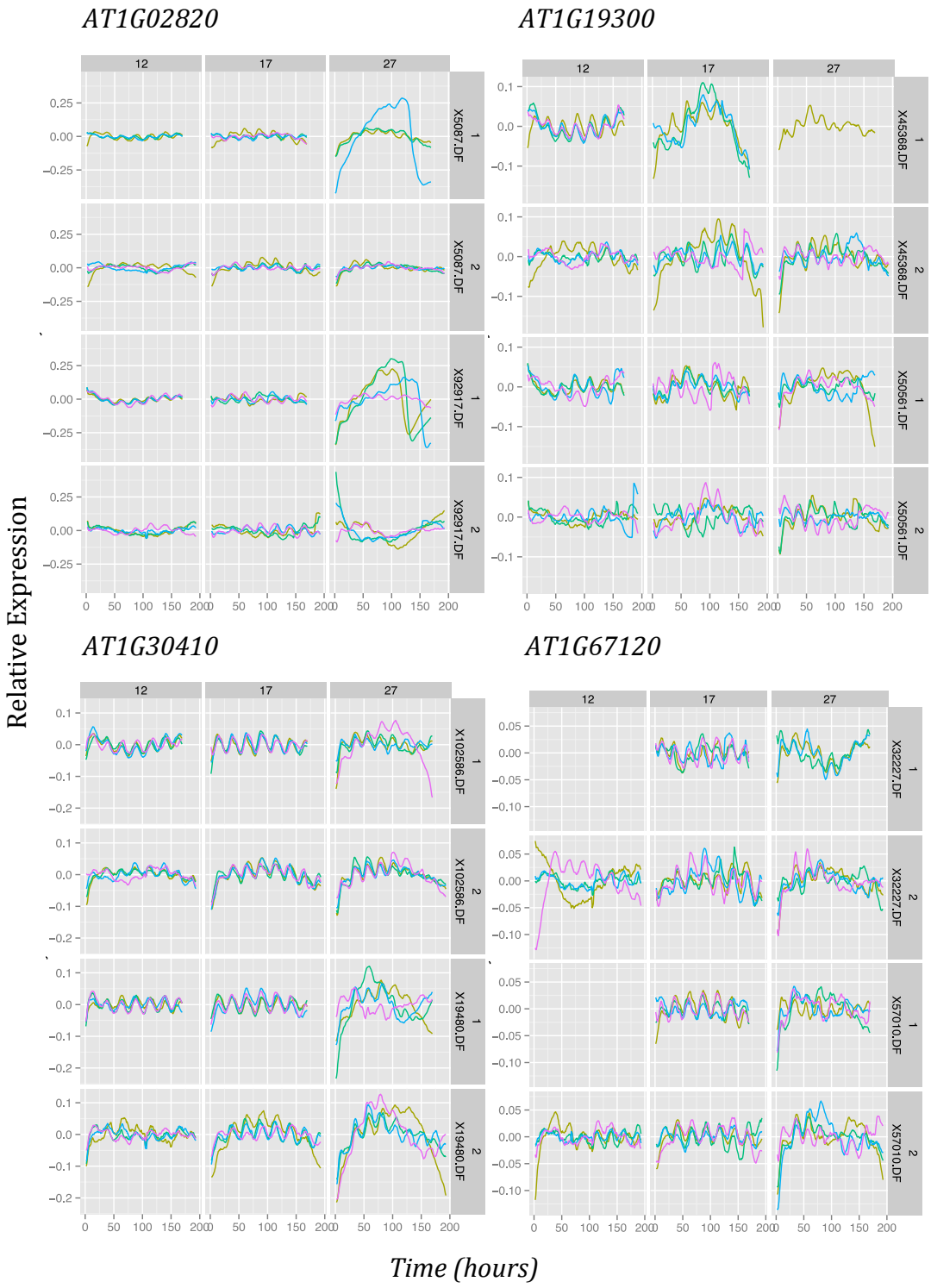
3.6.1 – Choosing Genes

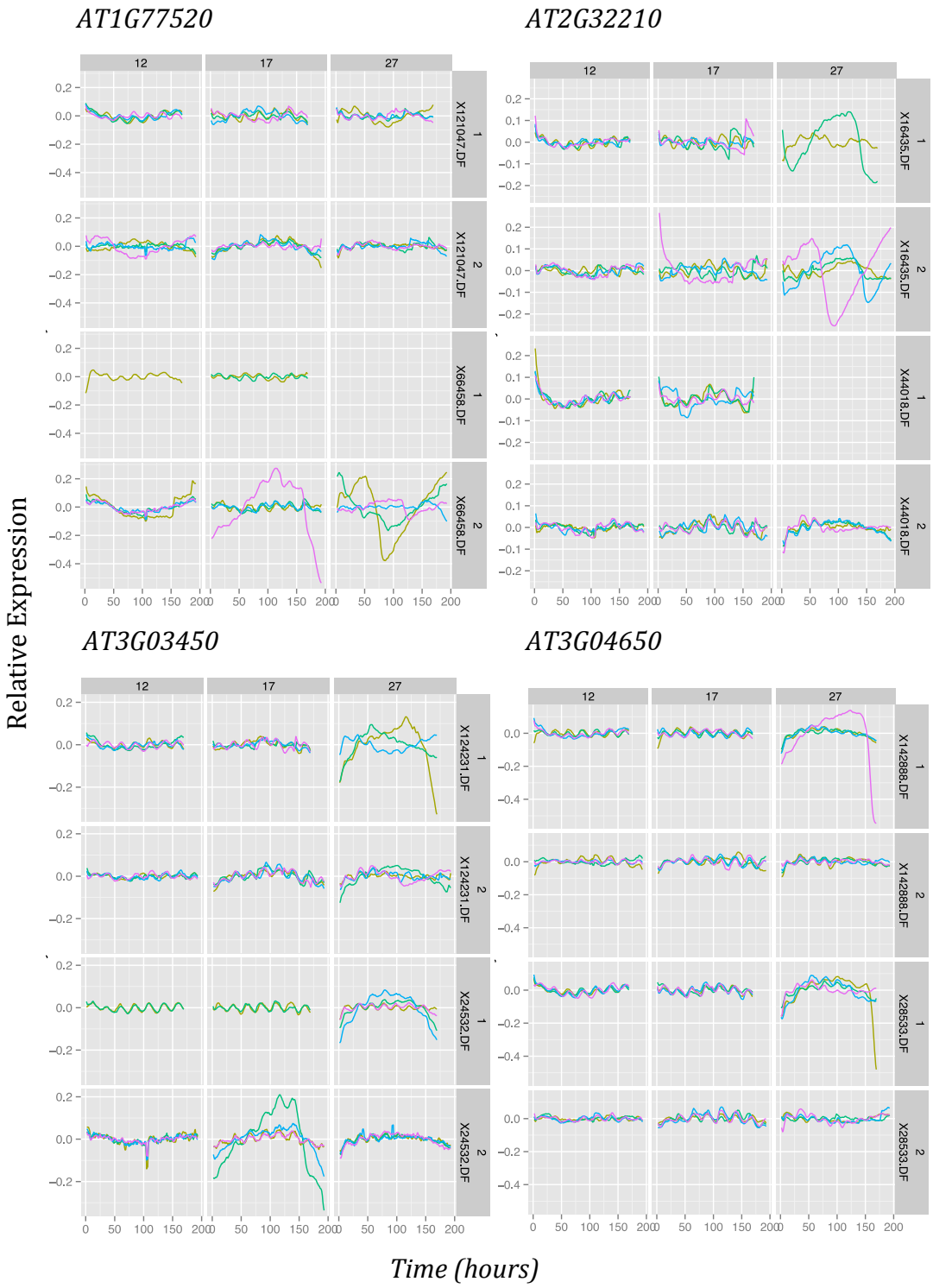
Considering fold changes between the microarrays created the list of temperature and GI sensitive genes. First, a gene needed to have a two-fold expression change from its lowest expression to its highest expression in the wild type plant (this was usually from 12˚C to 27˚C or vice versa, however there were a few exceptions). In addition, a gene had to have an effect in its expression between the wt and *gi-11* microarrays. This was achieved by scaling the fold change between wild type and the mutant at each temperature using a $\log_{1.5}$ function. If any single temperature had a value of 1 or -1 (i.e. the gene had a 1.5x fold change at any single temperature) it was classed as significantly changing its expression in the *gi-11* mutant. Similarly, if the maximum scaled fold change minus the minimum scaled fold change (again this was usually between 12 and 27˚C) had a value greater than 1, then the gene was considered to be sensitive to the mutant, but in a temperature dependent manner (Supplemental Table 3.4).

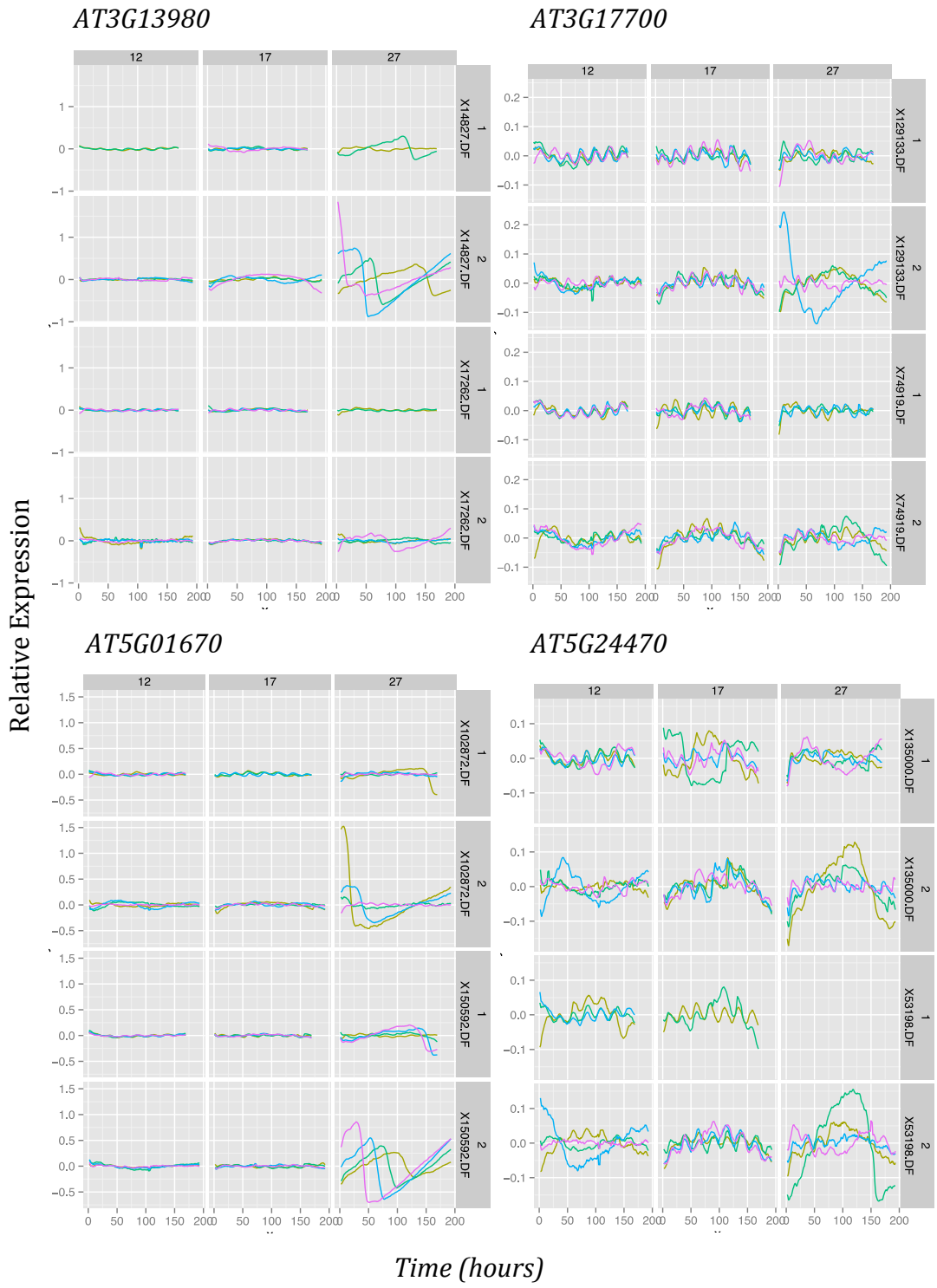
Using this list of genes, the SALK database was explored to identify genes that had two lines containing a homozygous mutation in either the promoter or exon regions. This resulted in 72 genes (144 SALK lines) that were ordered from NASC. The lines for four of these genes failed to germinate, leaving 68 genes to screen using DF. Of the genes screened, 5 produced data for only one line.

3.6.2 – Results

The remaining 68 genes were passed through DF experiments performed at 12, 17 and 27˚C in duplicate by Peter Gould. The remaining plants had their delayed fluorescence expression profiles analysed for a change in period length relative to wild type. This was done by first calculating each plant's circadian period using Spectral Resampling (Costa et al. 2013). Deviation from wild type plants was then examined using 2-sample t-tests (Supplemental Table 3.5). Due to the nature of the statistics, genotypes where none or only 1 set of plants had a detectable rhythm received a result identical to a genotype where there is no rhythm. As such, the profiles were graphed out and cross-referenced with the outcome of the t-tests to check whether an inferred arrhythmia was real or just low repeat error (Supplemental Figure 3.1). By then considering what happened in each of the two mutants and across the temperature series, 13 genes were identified as having a temperature dependent circadian effect (Fig 3.9). These 13 genes were then cross-referenced against the TAIR website (Rhee 2003). From this, the inclusion of PRR3 and PRR5 in this list was identified. These genes were known components of the core circadian clock genes in Arabidopsis and PRR5 had previously been identified as having a temperature dependent phenotype (Salome et al. 2005). Additionally, multiple genes involved in photosynthesis related processes were found. The effect of these genes on the DF experiment was likely complex, with some elements that affected the reporter system, and others that might have affected the circadian clock. Other genes identified, however, did not have an interesting annotation, and some did not have an annotation in the database at all. A summary of these genes and their function is shown in Table 3.2.


AT1G02820
AT1G19300
AT1G30410
AT1G67120
12
17
27
X5087.DF 1
X5087.DF 2
X92917.DF 1
X92917.DF 2
X45368.DF 1
X45368.DF 2
X50561.DF 1
X50561.DF 2
X102586.DF 1
X102586.DF 2
X19480.DF 1
X19480.DF 2
X32227.DF 1
X32227.DF 2
X57010.DF 1
X57010.DF 2
Relative Expression
Time (hours)


AT1G77520
AT2G32210
AT3G03450
AT3G04650
12
17
27
X121047.DF 1
X121047.DF 2
X66458.DF 1
X66458.DF 2
X16435.DF 1
X16435.DF 2
X44018.DF 1
X44018.DF 2
X124231.DF 1
X124231.DF 2
X24532.DF 1
X24532.DF 2
X142888.DF 1
X142888.DF 2
X28533.DF 1
X28533.DF 2
Relative Expression
Time (hours)


AT3G13980
AT3G17700
AT5G01670
AT5G24470
12
17
27
X14827.DF 1
X14827.DF 2
X17262.DF 1
X17262.DF 2
X129133.DF 1
X129133.DF 2
X74919.DF 1
X74919.DF 2
X102872.DF 1
X102872.DF 2
X150592.DF 1
X150592.DF 2
X135000.DF 1
X135000.DF 2
X53198.DF 1
X53198.DF 2
Relative Expression
Time (hours)

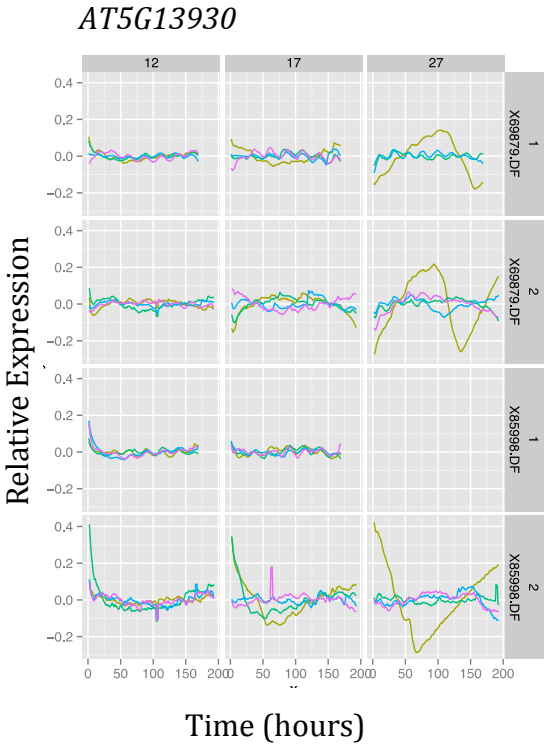


Figure 3.9 Delayed fluorescence time course of mutant knockouts. Each gene had two mutants (NASC ID's on the right) and was performed in duplicate. This was done at three temperatures (top) under red/blue light. Each line shows an individual set of plants recorded.

Table 3.2 Genes with a significant delayed fluorescence phenotype and their summary description

AT number	Gene name	Description
AT5G01670	NAD(P)linked oxidoreductase	involved in oxidation and reduction, located in cytoplasm
AT3G13980	unknown	located in nucleus
AT3G17700	CNGC20	involved in cyclic nucleotide binding and ion channel activity
AT1G30410	MRP13	involved in transmembrane transport
AT1G67120	MDN1	Yeast homolog possibly involved in ribosomal assembly
AT1G77520	O-methyl transferase	involved in lignin biosynthesis, located in cytosol & nucleus
AT1G02820	LEA3	involved in embryo development and response to stress
AT1G19300	GATL1	involved in xylan biosynthesis
AT3G03450	RGL2	involved in flower and fruit development, located in nucleus
AT3G04650	FAD/NAD(P) binding	located in chloroplasts and mitochondria
AT2G32210	unknown	unknown
AT5G60100	PRR3	involved in circadian rhythms, located in nucleus
AT5G24470	PRR5	involved in circadian rhythms, located in nucleus

3.7 – Discussion

Within this chapter, several ways by which plants compensate their biology for varying ambient temperatures were identified. Firstly, it was found that whilst this temperature range did not exceed the plant's natural conditions, stress responses were starting to activate to deal with the temperature. A significant number of genes involved in the cell wall biochemistry became differentially expressed with changing temperature, suggesting that plant growth occurs at different rates at different temperatures. However, this response in growth was not as simple as genes being up regulated in high temperatures, many genes were also up regulated at lower temperatures. Additionally, it was found that the way that genes were regulated by transcription, translation and degradation changed across the temperature range. This appeared to be occurring in a balanced manner. The transcription factor activity and protein degradation pathways were up regulated at high temperatures. However, the plastidic translation of mRNA was up regulated at low temperatures. This was supported by the idea that changes in transcription and translation rates are drivers of general temperature buffering of the organism (Sidaway-Lee et al. 2013).

However, the hypothesis that the buffering of the circadian clock controls how the plant responds to temperature changes was not supported by this data. When the circadian clock had it's ability to adapt to temperature impeded, in this case through the mutation of GI, there were no visible wide spread effects on the general response of the plant to temperature. The fact that the circadian process was reported to change in this mutant (Gould et al. 2006) suggested that the clock's adaptation to temperature is not caused through generic changes to plant biochemistry, and that these two mechanisms are distinct from each other. Examining individual components of the core circadian clock in both wt and *gi-11* also supported this idea that a GI mutant causes temperature sensitivity in the clock.

In addition, 68 genes with a strong response to temperature and the *gi-11* mutation were screened to identify whether they had an effect on the circadian

clock. This screen was able to identify 13 genes with a significant effect on the circadian clock. Many of these mutant effects also exhibited a temperature dependent phenotype, suggesting that their roles may either directly or indirectly affect the temperature compensation efficiency of the circadian clock pathways. This included known clock components PRR3 and PRR5, which had previously been shown to have a minor role in sensing thermocycles (Eckardt 2005). These genes do not currently exist in circadian clock models, except through a hypothetical component NI. These genes were the only ones discovered that were already annotated as core circadian clock genes. As such, they are likely to be major factors needed to model temperature compensation, specifically in the previously identified GI pathway.

Several of the genes identified for the DF screen have previously been identified as being located in chloroplasts (Bayer et al. 2011). These genes may provide information as to how the circadian clock is coupled to photosystem II, a key criticism of the delayed fluorescent screen (personal communication, James Hartwell, Anthony Hall). Many of the genes with significant DF phenotypes have also been reported as genes involved in salt stress (Li et al. 2013; He et al. 2005). Salt stress has previously been identified as being involved with temperature compensation (Fry 1958), as well as the photosynthetic pathway (Downton et al. n.d.). This suggests yet another way that the circadian clock may be connected to the photosynthetic pathway, in addition to how the circadian clock is buffered against ambient temperature change.

However, this investigation had some limitations. Firstly, not all of the identified temperature and GI responsive genes had a SALK T-DNA tagged knockout line associated with them, creating the possibility that a major gene may not have been screened here. It was also still not fully understood how PSII connected to the clock, so it was possible that there was missing information due to an improper reporter system. In addition, this study was performed only on the transcriptome. This removes any translational, post-translational protein modification or splice variation that could be occurring (James et al. 2012; O'Neill et al. 2011). Additionally, the pooling strategy removes the ability to

monitor period differences. This prevented the investigation into temperature compensation, which is defined as free running period differences. With advances in microarray technology, this study can now be repeated. In this repeat, the individual samples can be used for microarray analysis instead of a 24 hour pooled sample.

Chapter 4 – Clustering of Luciferase Data Identifies Co-Regulation of Genes in a Temperature Dependent Manner

4.1 – Introduction

It has been previously shown that temperature compensation within the circadian clock was not linked to generic responses to temperature by the plant (Chapter 3). As such, it became interesting to investigate how the circadian network is temperature compensated if it is not driven by a plant wide temperature response. A key question was whether the network uses several toggle switches, with splice variants or homologues being used at different temperatures, or whether the entire topology of the network changes. Toggle switches would have caused only the genes in the switch to have an altered expression relative to the rest of the genome. However a topological change would have caused far greater changes in gene expression relative to the other genes tested.

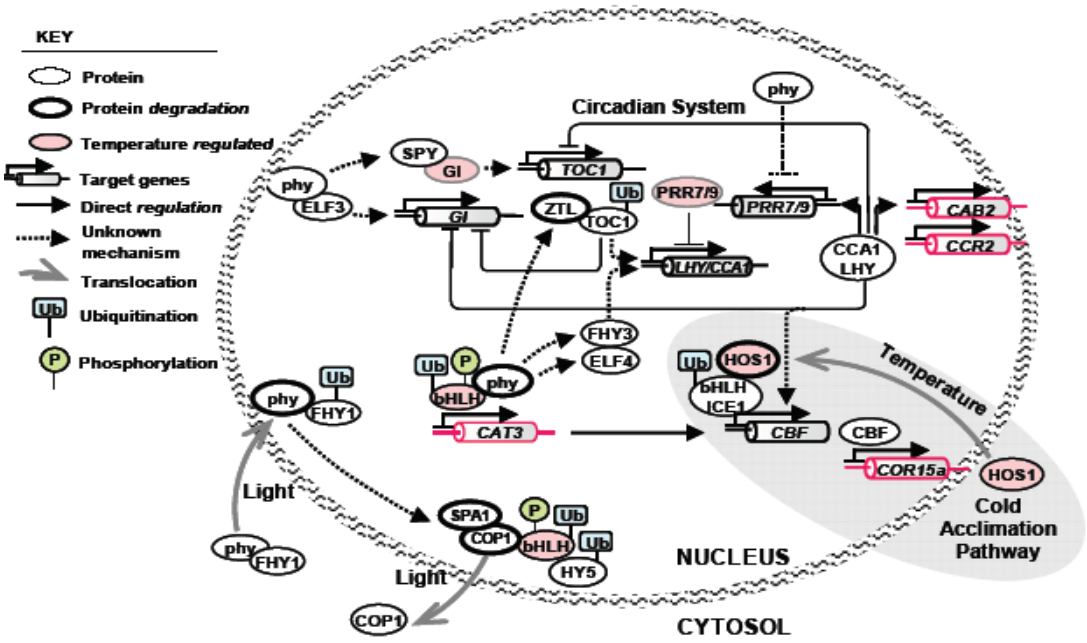
To investigate whether the topology of the circadian network changed with temperature, gene expression profiles at a range of temperatures were gathered. However, as data generation becomes faster, and with a greater throughput, methods to quickly deal with the expanding data sizes became more important. A good first step was to reduce data quantity by grouping similar sets of data. This idea was called clustering and could be implemented in numerous ways (Jain et al. 1999). The essence of most methods was to form multiple small groups, each containing a subset of the data that are ranked as most similar to each other, based on various defined criteria. In this section, clustering methods were used to analyse high-resolution promoter::luciferase time courses produced under a number of experimental conditions. These time courses consisted of promoter linked luciferase constructs, grown at 22˚C for 7 days in LD cycles. These were then moved to imaging chambers and experimental temperature. They were then recorded for 2 days under LD cycles followed by 5 days in constant light (Gould et al. 2006). If the network did indeed change, these clusters would change depending on the conditions used to generate the data.

Once clusters were produced, the next step was to understand how these clusters changed with temperature. To do this, a method was developed which compared the clusters produced under each of the conditions. This method was then evolved to provide a visualisation that showed each of the cluster sets produced as well as information on genes that were clustered together under multiple conditions. Using this visualisation tool, several known co-expressed genes (e.g. LHY and CCA1 Gould et al. 2006) were identified as being clustered together under all the conditions. Whilst this could have been identified from the raw data, the visualization tool allowed for additional complexes to be identified (e.g. ELF3 and SPA3). However, it was far more common for gene clusters to vary across the different temperatures. This supported the hypothesis that the network topology was changing with temperature. This change in network topology would cause the similarity in gene expressions to vary over this temperature range.

4.2 – Data Origins

For a single set of conditions, the data consisted of 42 different promoter::LUC constructs. These existed in multiple lines, which represented different transformation events. These lines ensured resulting data was not a product of where in the genome the insertion took place. The promoters chosen were all associated with genes that were thought to be part of the circadian clock mechanism. These genes might be part of the core oscillation circuit (e.g. LHY or TOC1), part of signaling input (e.g. PHY's or CRY's), or part of the output pathway (e.g. CAB2 or CCR2). (Fig 3.1 A) These were then screened using the method detailed in Chapter 2.2.4.2. This experimental setup was run under a range of light and temperature conditions. In each set of conditions, the overall method was the same, i.e. constant temperature and 2 days LD followed by 5 days LL. However the light used could either be 40 μmol/m^2/s of blue light, 40 μmol/m^2/s of red light or 20 μmol/m^2/s of both red and blue light. Additionally the plants were grown at 12, 17, 22 or 27˚C (Figure 4.1 B). These conditions were done in combination giving a total of 12 different data sets. Each set was checked for anomalies (e.g. 'hot pixels', produced due to electrical noise) before being inputted to BioDare. Missing data due to plants dying was replaced using averages of the remaining data series. This was done in such a way that no individual plant or line received preferential consideration within the full data set.

A)



B)

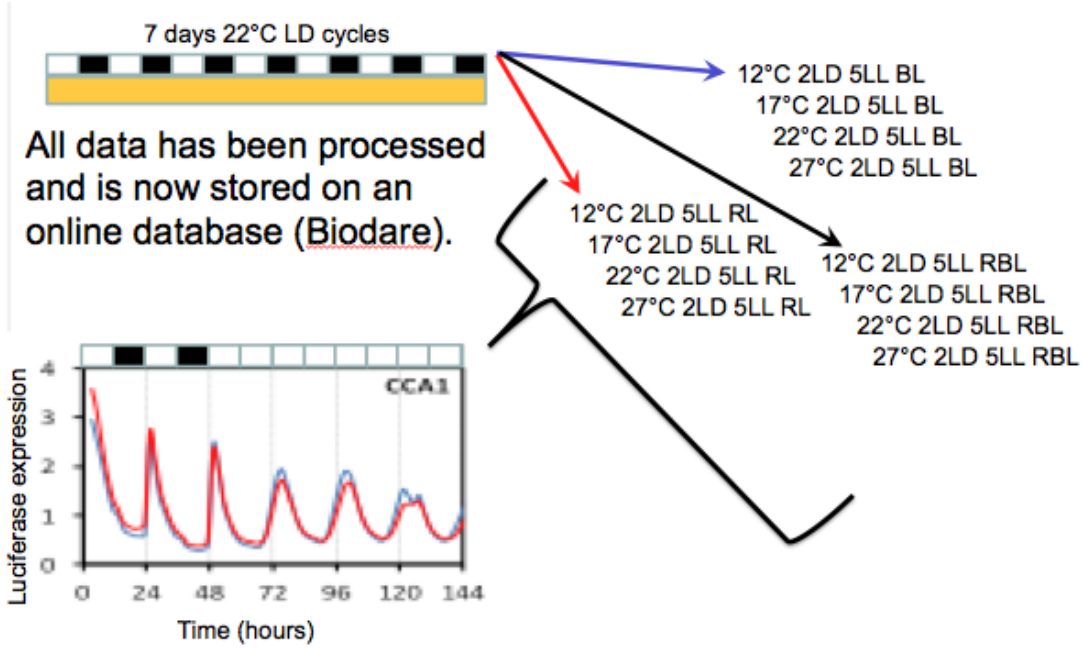


Figure 4.1 Gene selection and experimental design of luciferase data. A) Genes selected for promoter::LUC constructs and how they fit into the circadian network. B) Experimental set up of luciferase screens. Plants are grown for 7 days in LD cycles at 22˚C before being moved to experimental light and temperature conditions. These are then imaged every 2 hours for 2 days LD and 5 days LL.

4.3 – Clustering Methods

There were many ways to cluster data sets, each with it's own strengths and weaknesses (Jain et al. 1999). A major division in cluster software was how they determined how genes fitted into clusters with each repeat of the algorithm. A mid-point cluster algorithm created a set of random cluster points and then placed each datum into the best fitting group (based on distance to the cluster point). The cluster point of each group is then updated to match the average of the data in that group. Each datum is then reanalysed and, if needed, placed into a better fitting group. This was repeated until no data changes group, or for a maximum number of iterations. The results from this method were highly dependent on the starting cluster points, and running several times was likely to generate different results. There was also the complication of choosing the number of clusters you expect to have at the end, especially in non-trivial data sets.

The alternative method was termed hierarchical clustering. In this method every data point started as its own cluster. The two most similar clusters (as calculated by specifics of the algorithm, usually a distance score) were then joined together to form a new single cluster, which was defined by a summary pseudo data point of all its members. The next two most similar clusters were then joint etc. until every point was within the same cluster. A 'best fit' cluster set could then derived based on number of clusters, the within and between standard deviations, as well as a probabilistic likelihood statistic.

Additional to the method of clustering, decisions needed to be made on how the data should be interpreted. Some software packages will deal with multiple values per data point by considering it to be a point in an n-dimensional space. In these packages, the order of the different values did not affect the clustering results. In contrast to this, some software packages treated the order of values as importantly or even more important than the actual values.

The data being analysed were specifically time courses of known oscillatory genes, with proven differences in peak timing and period lengths. As such, it made more sense to utilise a software package that considered each series as a series rather than a multi dimensional object. Also, given the complexity of the curves, using a hierarchical approach was more intuitive since it would be difficult to assess how many clusters would be appropriate. Given this, two potential software packages to use in clustering our data were identified, SplineCluster (Heard et al. 2006) and FFTSpline (Liverani et al. 2009).

4.3.1 – Pre-processing Data

Data coming off Biodare was luciferase data that had been normalised and detrended. However this meant it was potentially a luciferase affect that was being analysed, which might have obscured key features or introduced anomalies. To undo this effect, a back calculation was required to return to a time series that more represented the genes transcription rate. This was done using another software developed by collaborators on the ROBuST project namely ReTrOS (Costa et al. 2014). Using experimentally generated values for the translation of LUC mRNA, this software corrects the timing and amplitudes of peaks in the luciferase expression time series. The effect of this back calculation could be seen in the representative graph of CCA1 seen in fig 4.2. Additionally, the luciferase reporter enzyme was discovered to have effects that were temperature dependent. This led to different amounts of light being produced at each temperature. By performing this back calculation on each of the individual temperature conditions, the data produced by ReTrOS also made the results generated at different temperatures more comparable.

In addition to removing luciferase effects, data was investigated for whether it needed to be split into the two stages that occurred in the experiment, namely LD and LL. Light driven cycles (LD) had stronger, higher amplitude, and more robust oscillations. These were also more conserved across the different genes with multiple genes having a very similar LD time series. In LL, however, the

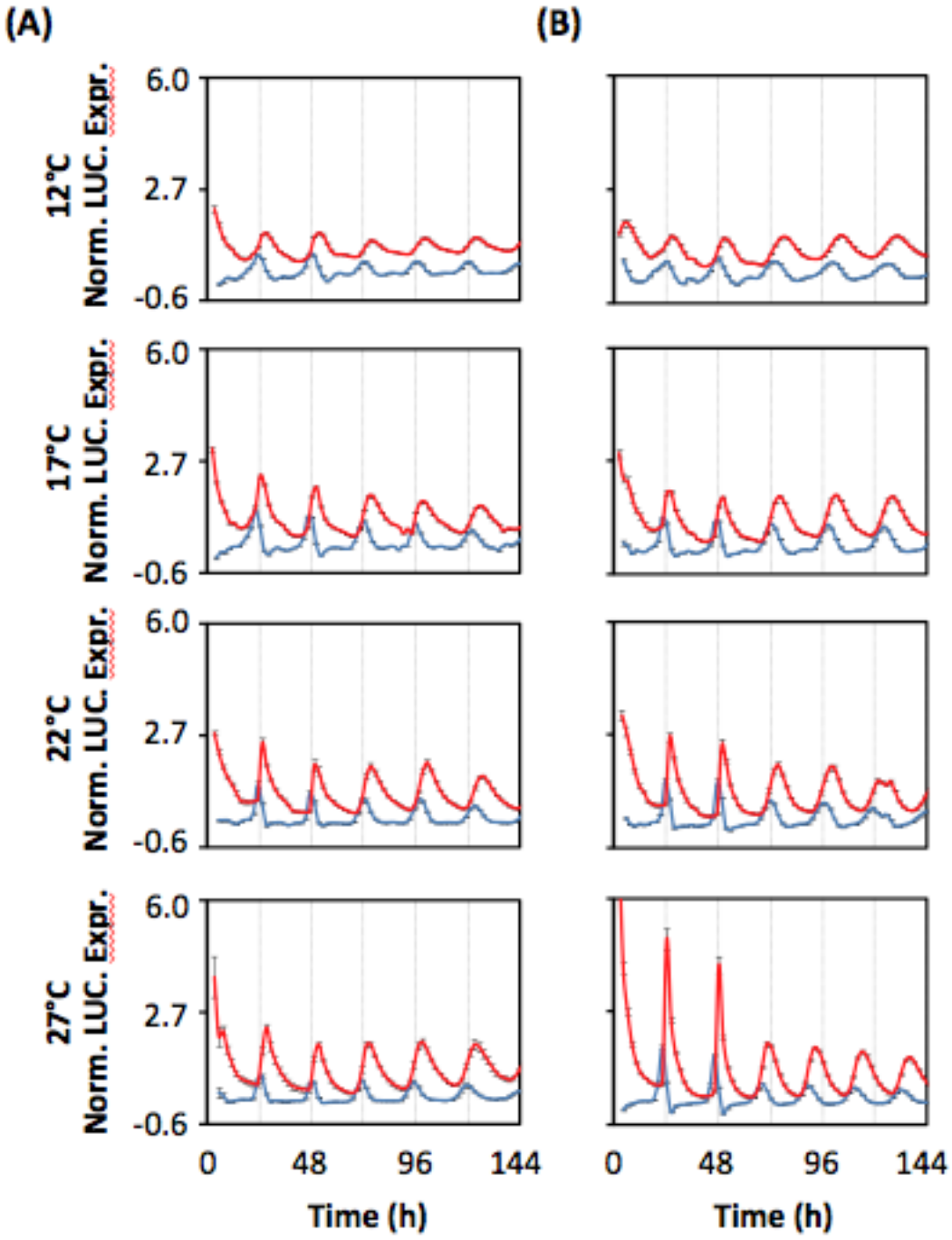


Figure 4.2 Effects of ReTrOS on CCA1. Raw luciferase data (red) and ReTrOS corrected data (blue) for the different temperatures as well as for A) red light and B) blue light.

effects of dampening became an important difference between genes. However some genes became completely arrhythmic soon after entering constant light. By comparing results generated from the three possible light regimes (LD, LL and LDLL), genes can be investigated by how they behave relative to each other not only in driven cycles, but when they were able to free-run.

4.3.2 – Software Comparison

Both SplineCluster and FFTSpline used splines fitted onto curves to summerise complex waves forms into a set of principal characteristics. It was then these characteristics that were clustered by the hierarchical clustering algorithm to form cluster lists. SplineCluster took each of the raw data series and fitted beta-splines onto it. This was done dynamically, with sections of the time series with great variability having move splines, and sections with low variability having very few splines. This resulted in characteristics such as the acute induction of LHY around dawn being accurately represented. FFTSpline uses the same method of fitting beta-splines, however performs a FFT (Fast Fourier Transform (Plautz et al. 1997)) algorithm to smooth the curve before fitting splines. FFT has been widely used in oscillatory systems for many years and provided a quick method of period estimation. Consequently, it was investigated whether its use might have provided better data for clustering by removing noise in the oscillations from the data. This was likely to be especially important in the LL data set, where acute induction was not present. However, when applied to assymetric curves, a FFT transformation risks altering the peaks of the data. When clustering the data set, the FFTSpline software frequently produced clusters where genes with anti-phase curves were being clustered together (Fig 4.3). In cluster 2 of Fig 4.3, over half the genes were clustered together despite several distinct oscillations being visible. Similarly, in cluster 3, one gene appeared to completely out of phase of the other two. Additionally, the blue line in each cluster represented the fitted average to that cluster. Throughout these were noisy, but it was most striking in cluster 4, where the data looks fairly close to a nice, strong oscillation, but the fitted curve does not match this. This

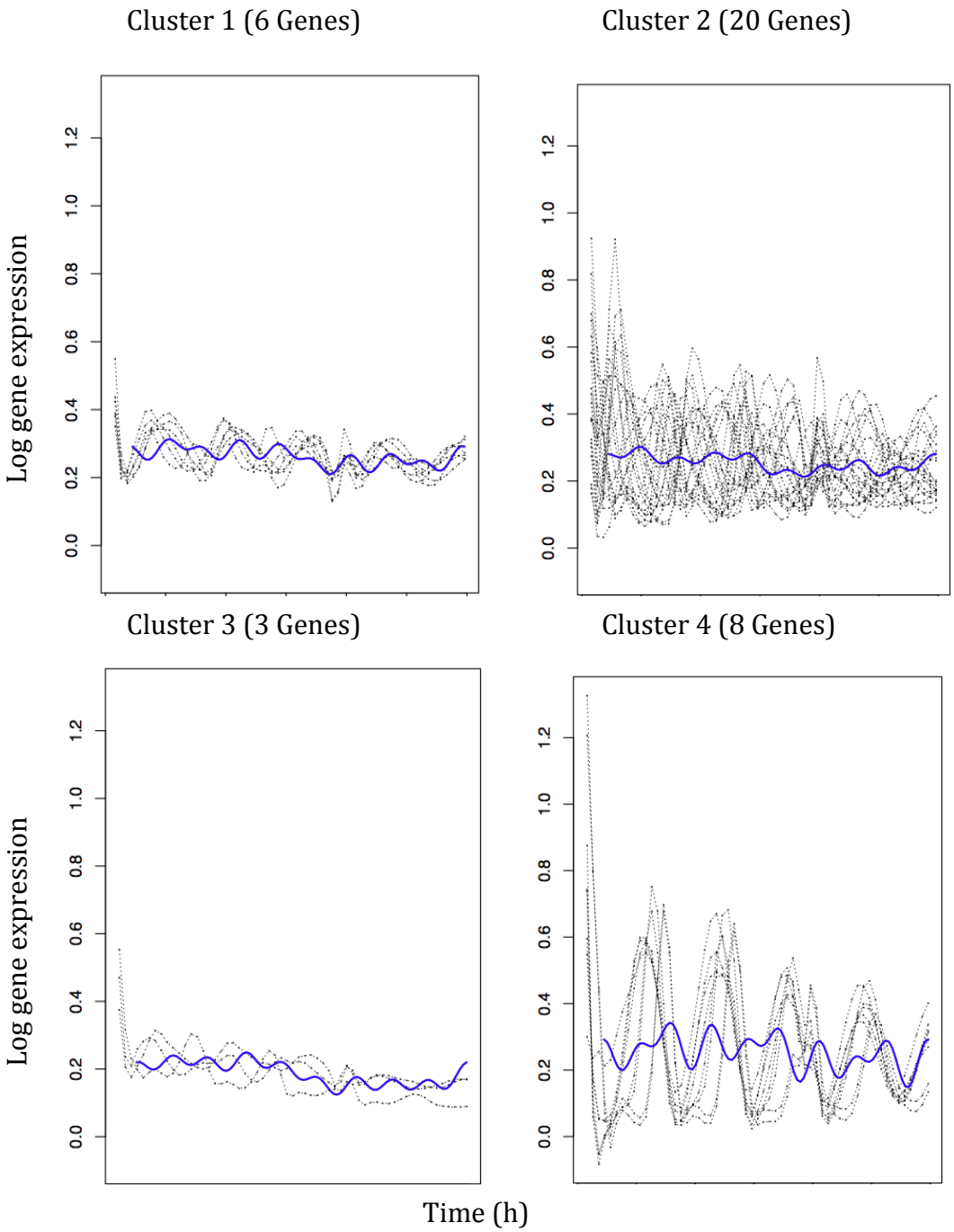


Figure 4.3 FFT-Spline cluster results of plants grown in blue light at 22˚C. These four clusters have multiple concerns. The second cluster includes a lot of rhythmic genes of different phases, whilst the third cluster contains only 3 genes but which appear to be doing different things. The blue line shows the fitted average whilst the black lines show individual trends.

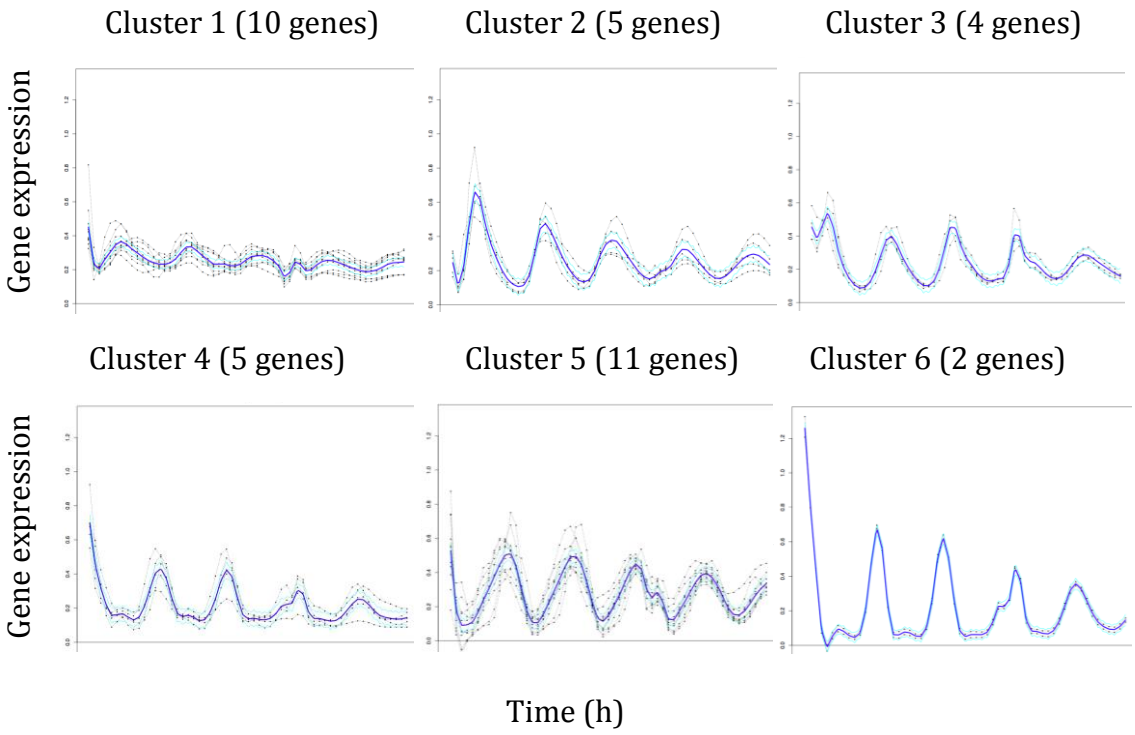


Figure 4.4 SplineCluster cluster results of plants grown in blue light at 22˚C. These six clusters all have nice consistency with only the odd point here or there that doesn't match. The blue line shows the fitted average whilst the black lines show individual trends.

false clustering was likely due to the non-sinusoidal nature of circadian oscillations. Using these clusters did not produce reliable results, and the large number in cluster 2 made further analysis difficult as so many things were clustered together without actually being similar. In contrast to this, results from SplineCluster showed no obvious anti-phase combinations (Fig 4.4). Additionally all fitted lines closely matched the raw data. Both of these cluster sets were produced from constant light for blue light 22 ˚C, using the average expression of each gene.

4.3.3 – Clustering Results

Clustering was done independently for each of the two light regimes (LD and LL) as well as for the combined LDLL data. It was also done for each of the three light conditions (red, blue, and red/blue), as well as the four temperatures (12, 17, 22 and 27˚C). Clustering the individual profiles usually resulted in only one or two different genes occurring in each cluster, with the gene repeats forming the bulk of what was clustered together. Many genes did show some variation in how they clustered, with a few repeats going into one cluster and others going into a different cluster (Supplemental Table 4.1), however these were usually closely related clusters, and the slight deviations were usually based upon which mutant line was used. To gain an understanding about which genes were being co-expressed in a temperature dependent manner, it was more useful to consider scenarios where the different genes form the basis of the clusters. This could have been done by clustering the clusters, or considering a higher threshold on the hierarchical tree.

However, a more practical solution was to cluster a gene's average profile. This resulted in more meaningful clusters (Supplemental Table 4.2). Unfortunately genes were not always present within every experiment. Major circadian clock genes LUX and PRR7 were found to be missing within several temperatures in the red/blue dataset. Conversely, FHY1, FHY1LIKE, HY5 and PHY were only present in red/blue light datasets. Additionally, GAI, PHYD and RGL3 were

missing from many of the datasets. Due to the central role of both LUX and PRR7, it was decided to focus on the red and blue datasets and ignore the red/blue dataset until data was gathered for the missing components. Clustering was rerun on the datasets that had extra genes, after first removing the additional information from the dataset.

As expected, when only the data gathered in LD conditions is clustered very few clusters are returned by the software, usually 2 or 3. Comparing results for clustering performed with LL and LDLL datasets, there were many similarities between the number of clusters, as well as the members of each of the clusters. Looking at gene combinations with known interactions/co-expressions, a lot of combinations that validated the method could be identified. For example, LHY and CCA1 were co-clustered in every cluster set. Since these genes were usually modeled as a single entity because they had such similar expression profiles, this was unsurprising. Similarly, the two phytochromes (PHYA and PHYC) left in the dataset were found to be co-clustered in all red light cluster results, but under blue light conditions they were sometimes placed in different clusters. Examining the table of clusters in this way was useful for examining sets of genes expected to be co-expressed for temperature and light effects, however examining this table became more difficult when attempting to see how two genes were clustered when little was known about their interaction with each other.

4.4 – Consensus Clustering

Clustering the gene time courses under each condition provided a good overview to what was happening at a global level, but it was difficult to see which genes were being clustered together across the condition sets. There were several software packages that would simultaneously cluster genes based on multiple sets of data. However, these methods tended to be a bit too stringent when used on these datasets, usually resulting in each gene being clustered on its own. Additionally, it was not only genes that were constantly co-expressed that were interesting, but also those that were only clustered together under certain conditions. Identifying this later type of cluster scenario was not readily available using current software. Additionally, due to the arbitrary nature of cluster labels, applying an automatic association by looking for genes in cluster x in each conditions was uninformative.

A solution to these problems was to analyse the cluster sets generated under each of the conditions and observe which genes remained clustered together. This was done best by converting the cluster list to a square matrix with binary elements. If two genes were clustered together, their intersects had a value of 1, if they were not, the intersects had a value of 0. To then compare two cluster sets, their respective matrices could be multiplied. The output showed those elements that were clustered under both conditions. Plotting this on top of an image showing how genes were clustered in one of the original conditions also showed which elements were clustered in that first condition, but not the second. Similarly, plotting the output of the matrix multiplication on the second cluster set showed elements that clustered in the second condition, but not in the first. Using a simple example, this idea could be displayed in a format that could be easily modified. To do this, a simple four-gene example that had been ‘clustered’ under three different conditions was produced (Table 4.1).

To maintain simplicity, only the first two condition sets were considered, and a tool that could quickly compare these two conditions was developed. Initially the transformation to Boolean matrices and subsequent multiplication methods

described above were applied. The values in this matrix were then relabeled so that genes that were clustered together in both conditions got the same number (Table 4.2). These results could then be graphically displayed using another section of code. This code utilised the multiple layer graphic package for R, ggPlot2 (Wickham 2009). It was designed to produce each of the four potential graphs; the two original cluster sets – where genes within a cluster were plotted around the circumference of a circle, with a different circle per cluster – and the consensus cluster plotted as a second layer on top of each of the starting clusters – the point for each gene became coloured, and lines were drawn between genes that were consensually clustered (Figure 4.5).

The code was structured such that it could take any table where the first column gives identifiers of the elements being clustered, and every subsequent column is the result of clustering them in some way. It also utilises column headers within titles to inform what is being plotted within the figure. The code then converted the table into multiple matrices. These matrices were then systematically multiplied together to create a consensus cluster for each possible set of conditions. The next subscript in the code produced all the appropriate images, plotting every consensus on each of the original cluster sets that made up the consensus. This meant that for table 4.1, this code produced a 12-page pdf. Pages 1-3 showed how the genes were clustered in each of the three conditions. Pages 4-9 showed each of the three pairwise comparisons, plotted independently on top of each of the appropriate images from the single cluster diagrams. Pages 10-12 showed the 3-way comparison, plotted independently on all three conditions.

This example data set highlighted some of the weaknesses in this original strategy, however. In table 4.2, it was apparent that three genes were clustered together under condition 2. However, in Figure 4.5 C that information could not be seen. This could be resolved by also considering either Figure 4.5 B or 4.5 D, however, it would be advantageous if a single image could have displayed all of this information.

Table 4.1 Sample cluster results for a hypothetical set of genes. Each column represents a different set of conditions used to generate data and each row represents a single gene.

	Condition 1	Condition 2	Condition 3
Gene A	1	1	1
Gene B	1	2	1
Gene C	2	2	1
Gene D	2	2	2

Table 4.2 Sample cluster results for a hypothetical set of genes. Each column represents a different set of conditions used to generate data and each row represents a single gene. The last column represents a relabeled consensus cluster of conditions 1 and 2.

	Condition 1	Condition 2	Combined
Gene A	1	1	1
Gene B	1	2	2
Gene C	2	2	3
Gene D	2	2	3

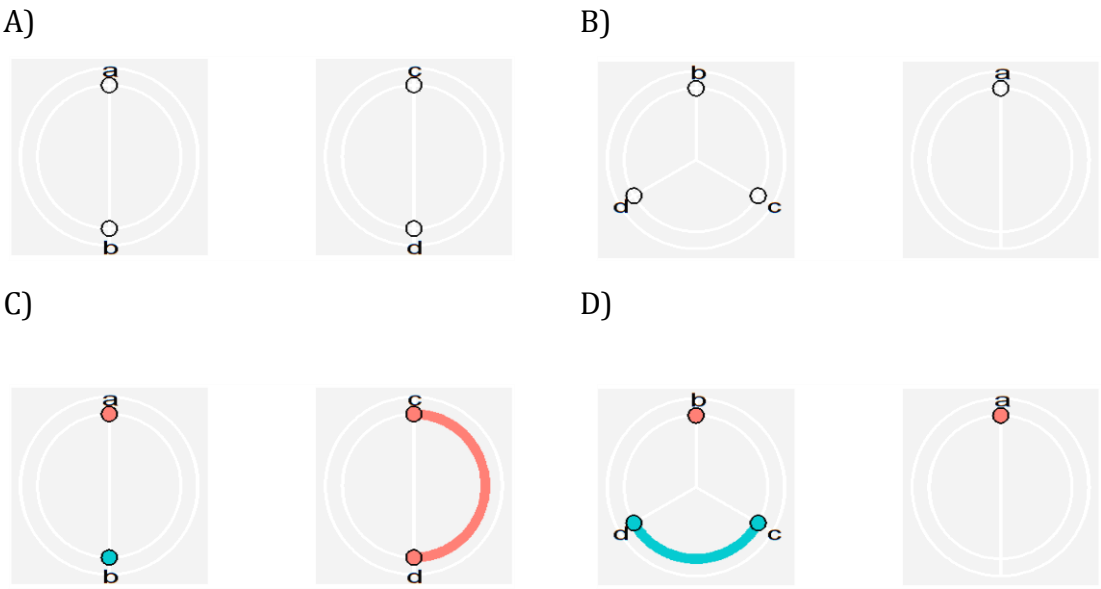


Figure 4.5 Cluster representations of Table 4.2. Figures showing how genes are clustered in our mock dataset. A) gene cluster membership in the first condition. B) gene cluster membership in the 2nd condition. C) consensus clustering in both the 1st and 2nd condition plotted on the backbone of the 1st cluster set. D) consensus clustering in both the 1st and 2nd condition plotted on the backbone of the 2nd cluster set. In A and B, the individual circles represented the different clusters and the white dots represented the genes within that cluster. In C and D, these dots were coloured depending on whether they were consensually clustered with other members of that cluster or not. Arcs were then drawn between genes that were consensually clustered. There was no meaning to colours between the different clusters, however.

This was an even greater concern when considering how the clustering varied as more conditions were used within the comparisons. Table 4.3 recreated table 4.1 but added the consensus cluster for the three conditions. As could be seen, the consensus for this set left every gene in its own cluster. Fig 4.6 displayed all three images that would be produced by the original code to describe the 3-way consensus cluster result. Very little information other than how the genes are clustered under the base condition could be identified from using these diagrams independently. Even when all three images were considered simultaneously, very little extra information could be recovered. Indeed using this method as it existed provided no easier method of considering the individual cluster results than already existed in the original table.

4.4.1 – Software Development

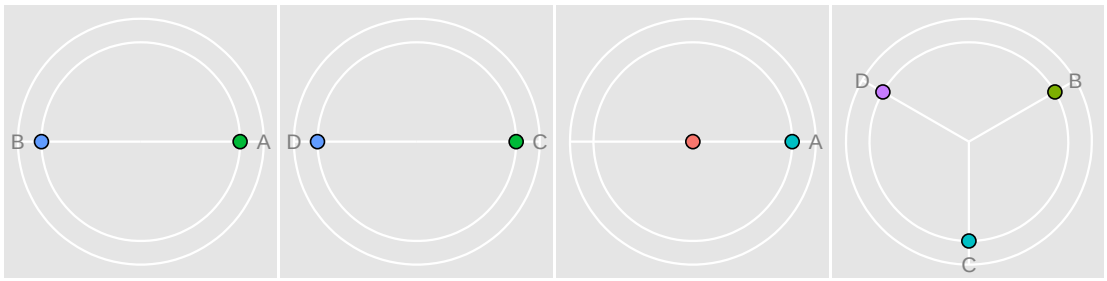
The first aspect developed on this software was how the graphical output represented two cluster conditions. Originally, the colour of the arcs drawn onto the base cluster had no encryption, they merely stated whether elements within each circle were clustered together in all conditions or not. There had been thoughts as to whether the colour should have been coded to the consensus cluster so that the same colours were not present on multiple circles but this proved difficult when the number of unique consensus clusters grew. However, in the simple case where there are only two conditions to display, it is instead possible that the colour could be used to indicate what cluster a gene was in when it was clustered by its expression in the second condition set.

Adding this feature to graphs composed of two clusters sets reveals more information than the original code. In fig 4.7, the consensus clusters for table 4.2 were redrawn to show this improved data retention. It was seen that these new figures made it clearer to see how gene B moved cluster between the conditions. With this additional level, not only can gene sets consistently co-clustered be identified, but also groups of genes that were clustered together in either

Table 4.3 Sample cluster results for a hypothetical set of genes. Each column represents a different set of conditions used to generate data and each row represents a single gene. The last column represents a relabeled consensus cluster of all three conditions.

	Condition 1	Condition 2	Condition 3	Combined
Gene A	1	1	1	1
Gene B	1	2	1	2
Gene C	2	2	1	3
Gene D	2	2	2	4

A) B)



C)

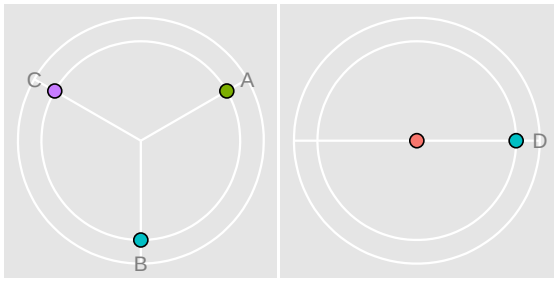


Figure 4.6 Consensus clustering output for a 3-way comparison. Figures were produced using the original consensus clustering code on table 4.3. The consensus clusters were added to the circles, which represented the cluster membership in: A) condition 1, B) condition 2, and C) condition 3. The individual circles represented the different clusters and the dots around the perimeter were coloured depending on whether they were consensually clustered with other members of that cluster or not. Arcs were then drawn between genes that were consensually clustered. There was no meaning to colours between the different clusters, however.

A)

B)

Figure 4.7 Consensus clustering of table 4.2. Circles represented the cluster allocation of condition 1 and spot/arc colours represented the cluster allocation of condition 2. A) diagram produced by the original code. B) diagram produced after colour encoding was introduced to the code. The individual circles represented the different clusters and the dots around the perimeter were coloured depending on whether they were consensually clustered with other members of that cluster or not. Arcs were then drawn between genes that were consensually clustered. In A there was no meaning to colours between the different clusters. In B, the colour of the dots can be directly compared between two clusters. Genes whose spots were the same colour were clustered together under the second condition.

condition but not both.

The next step was to develop the figures to allow a third condition to be visualised in a similar way. Several ideas were tested for how to incorporate a 3rd condition into this figure. The idea of concentric circles was initially tested, but when tested on a real data set it was realised that often genes could not be ordered in such a way that allows all genes that share a cluster to be put next to each other. The idea of using the arc to show one condition and the spot colour another was then considered to fix the above problem, however scenarios where a gene within the first cluster in the arc condition being the sole remaining member in the first cluster of the base condition would remove that information as there would be no second gene to draw the arc to (E.g. the left circle in Figure 4.7 B). Additionally an idea to use the shape of the point was rejected as it becomes difficult to distinguish shapes, especially if there are a lot of clusters or a large number of genes within a cluster making each point close to it's neighbours.

The solution to how to add a third data set came though utilising the radii of the circles. By drawing a radius from the centre of the circle to each point, there was an additional section that could be differentially coloured to provide added information. Figure 4.8 shows how this new 3-condition representation looked when it was applied to table 4.1. Only one of the possible outputs is shown, but the code produced all 6 possible ways of displaying this data. In contrast to figure 4.6, figure 4.8 contained all the information within a single image. It could now be seen that genes 1 and 4 were always different, and that genes 2 and 3 changed sequentially from being like 1, to being like 4. In this case, the extra information was trivial and could have easily been seen within the table, but when the number of genes and clusters increased, this visualisation became very useful. One important note, however, is that the colours used to draw the arcs and the radii were generated from the same list. However, there was no correlation between a radius having a specific colour and the arc being the same colour or a different one, just like a gene being clustered into cluster 1 in each condition had no added information. In addition, a forth layer was added to

Figure 4.8 Consensus clustering output for a 3-way comparison. Figures were produced using the updated consensus clustering code on table 4.3. The circles represented the cluster allocation in Condition 1. Colouring of the dots and connected arcs represented cluster allocation in Condition 2. Colouring of the radii represented cluster allocation in Condition 3. No additional information can be obtained by considering how the colour of the point compared to the colour of the radius.

these plots to allow all four temperatures to be visualised simultaneously. This was achieved by extending the radii out beyond the arcs. Now the circle a gene was plotted in was how it clustered in one cluster set, the colour of the point and any associated arcs was how the gene clustered in a second cluster set, the radius from the centre to the point was how it clustered in the third cluster set, and the radius extending past the point was how it clustered in the forth cluster set.

4.4.2 – Results

Using the visualisation method produced above, the clusters produced at each of the four temperatures can be compared. With this software now developed, the question became which light regime and colour was best to analyse. Producing the diagrams for blue light, the effect of the different light regimes could be explored (Fig 4.9, 4.10, 4.11). Comparing these images, it was seen that clusters produced using the full LDLL expression captured both elements seen in LD and those seen in LL. As such, using the clusters produced from the full LDLL time series was selected to better investigate how genes were co-expressed across the temperature range (Fig 4.11, 4.12). As previously noted, a strong connection is identified between LHY and CCA1 across all four temperatures and in both blue and red light. At 12˚C, they also clustered with other genes, many of which were conserved in both light conditions.

In addition to recovering known co-expressed combinations, several novel sets were also recovered. One example was ELF3 and SPA3, these genes clustered together in all eight conditions. Exploring what else they clustered with also produced interesting results. Under red light conditions, these two genes clustered with CCA1 and LHY at all temperatures, however this was only true at 12˚C in blue light. In contrast, under blue light they clustered with several different genes, which only appeared at 12˚C in red light. This shows that while they stayed very similar to each other across the conditions, depending on the light condition they were closer to different sets of genes at high temperatures.

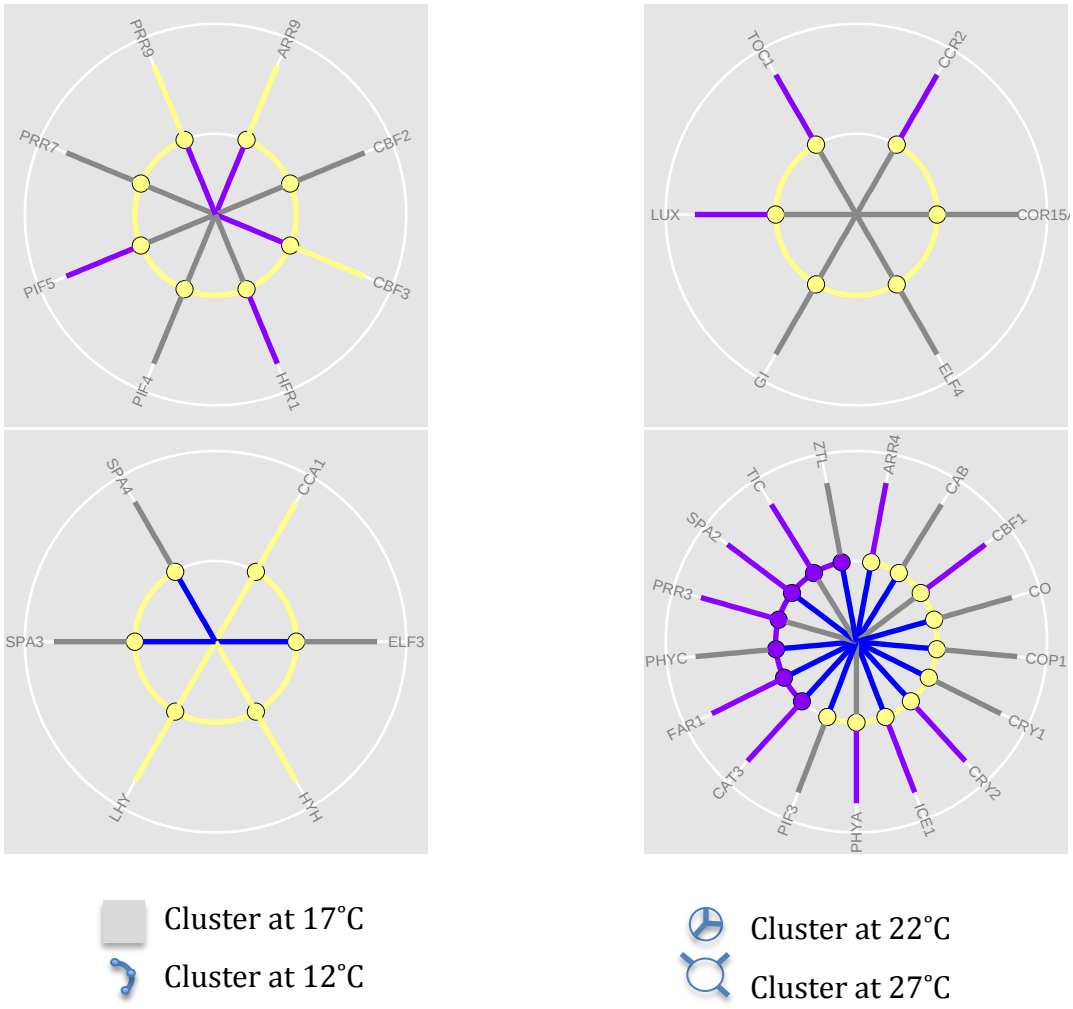


Figure 4.9 Consensus clustering of blue light data. Sample results for consensus clustering data produced in LD cycles. Circles represent individual clusters produced at 17˚C, spot and arc colours represent the cluster at 12˚C, inner radii represent the cluster at 22˚C and outer radii represent the colour at 27˚C. There is no connection between the colours of the different elements, however.

.

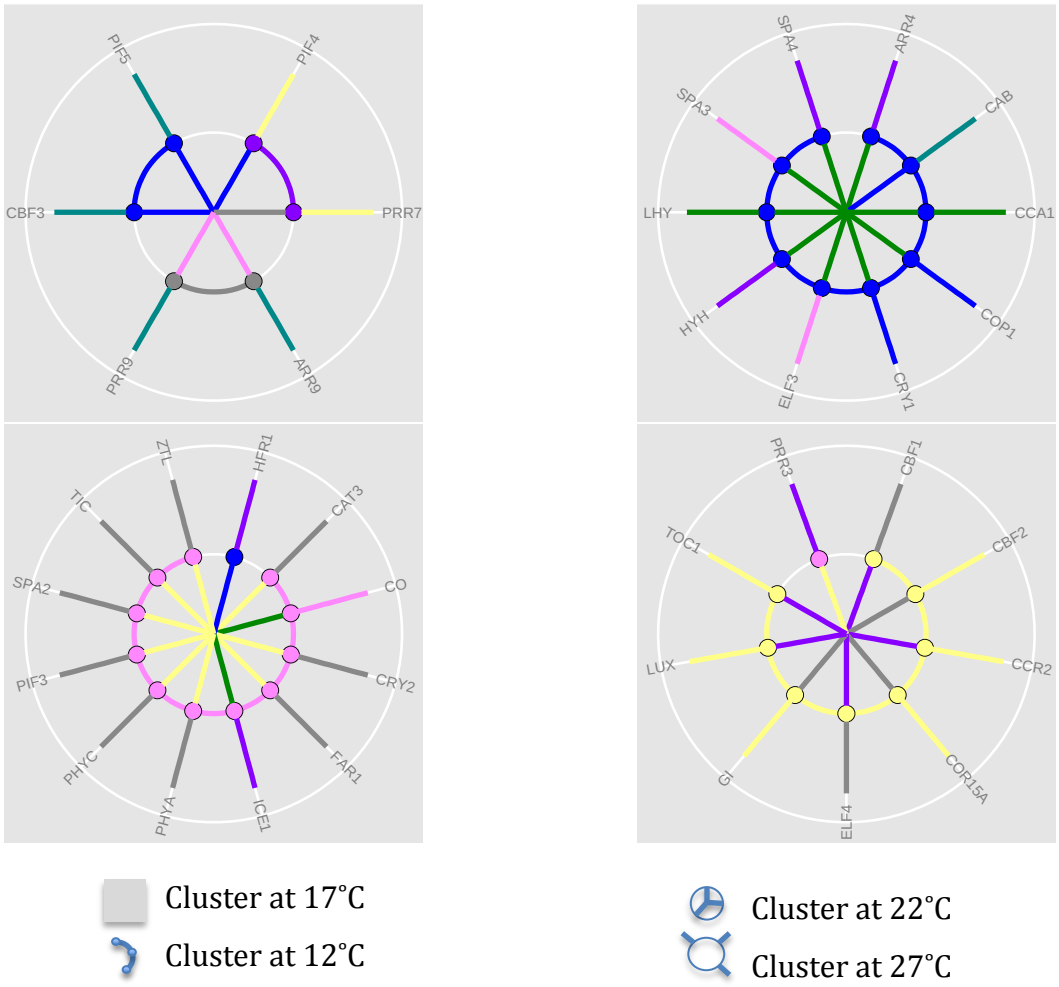


Figure 4.10 Consensus clustering of blue light data. Sample results for consensus clustering data produced in Ll cycles. Circles represent individual clusters produced at 17˚C, spot and arc colours represent the cluster at 12˚C, inner radii represent the cluster at 22˚C and outer radii represent the colour at 27˚C. There is no connection between the colours of the different elements, however.

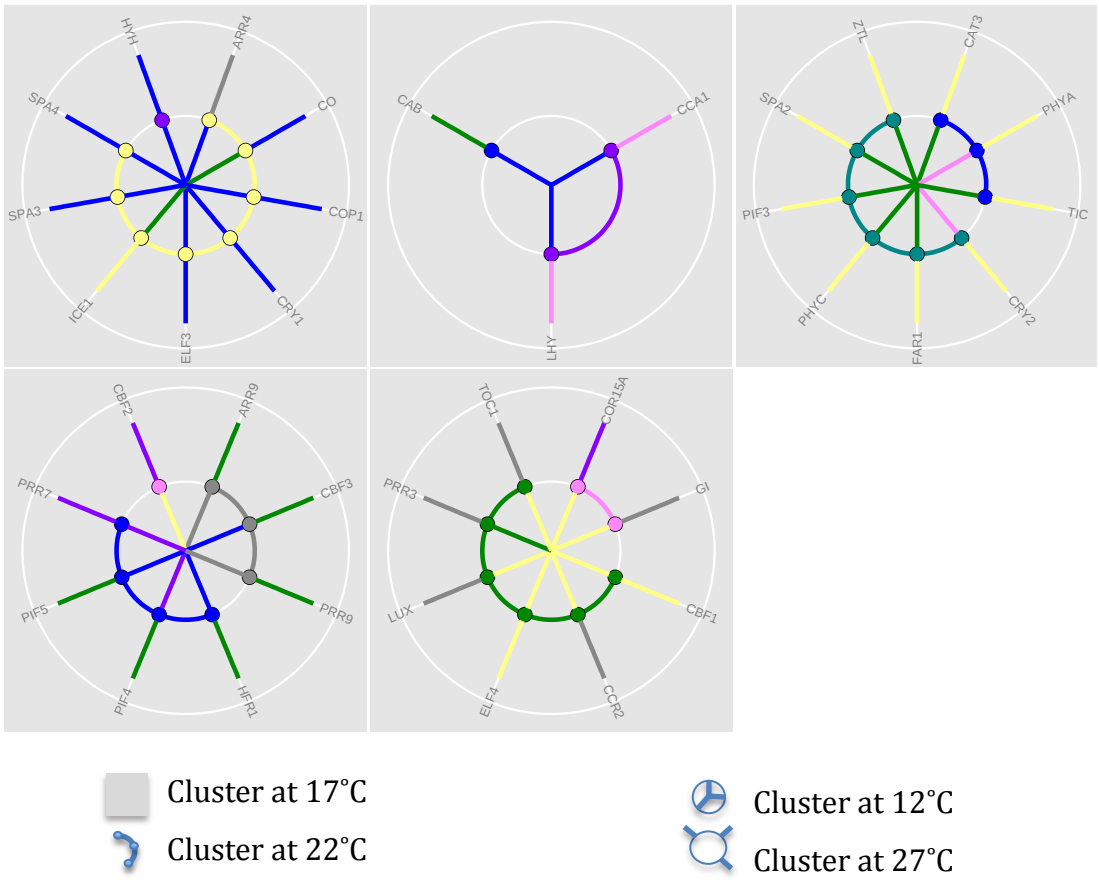


Figure 4.11 Consensus clustering of blue light data. Sample results for consensus clustering data produced in LDLL cycles. Circles represent individual clusters produced at 17˚C, spot and arc colours represent the cluster at 22˚C, inner radii represent the cluster at 12˚C and outer radii represent the colour at 27˚C. There was no connection between the colours of the different elements, however.

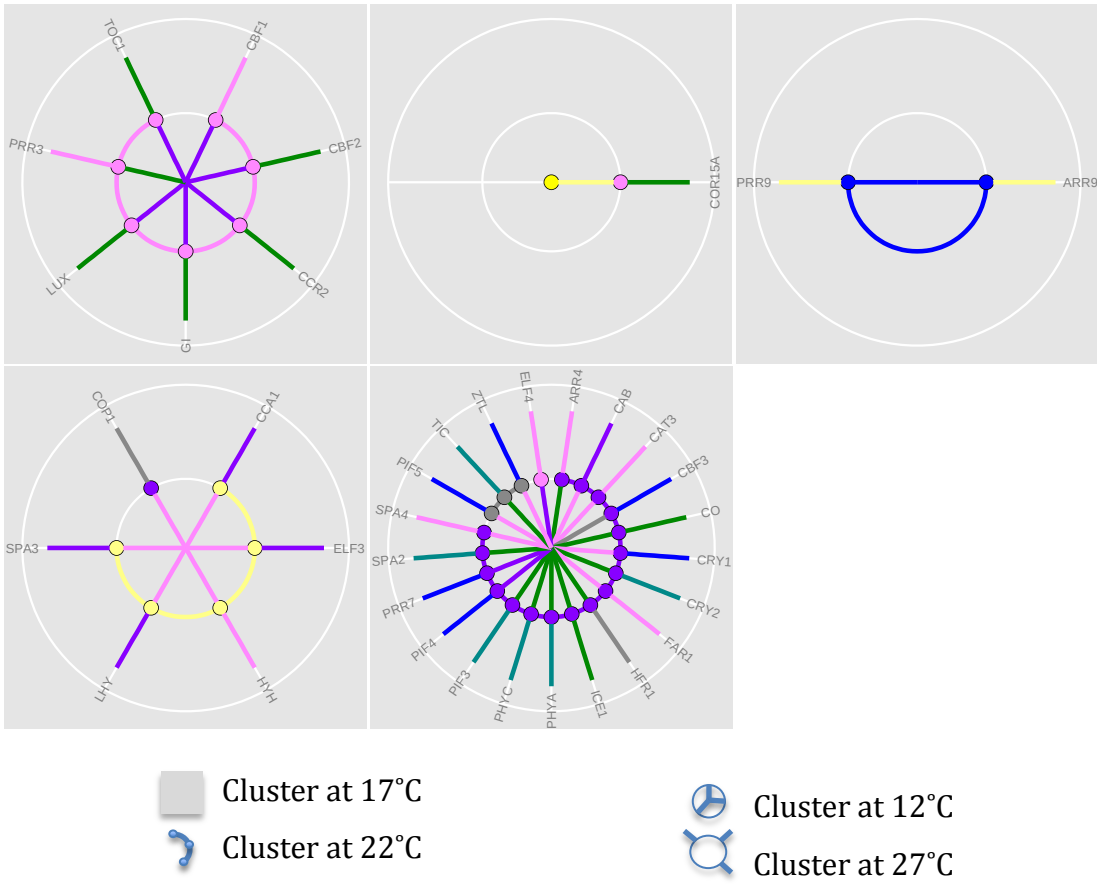


Figure 4.12 Consensus clustering of red light data. Sample results for consensus clustering data produced in LDLL cycles. Circles represent individual clusters produced at 17˚C, spot and arc colours represent the cluster at 22˚C, inner radii represent the cluster at 12˚C and outer radii represent the colour at 27˚C. There was no connection between the colours of the different elements, however.

To demonstrate how much information these diagrams showed, the first two circles from red light (Fig 4.12) were fully analysed. It was seen that CBF2, CCR2, GI, LUX and TOC1 clustered together under all conditions. CBF1 also clustered with these genes at 12, 17 and 22˚C, however did not at 27˚C. Similarly PRR3 did not cluster with these genes at 12˚C, but did at 17 and 22˚C, and then clustered with CBF1 only at 27˚C. This demonstrated how much can be recovered from a single circle, however, due to the encoding of the colours, more information could be recovered. The second circle showed that COR15A was clustered on its own at 17˚C. This gene did not cluster with any of the genes in the first circle at 12˚C, but clustered with all the genes in the first circle at 22˚C. At 27˚C it then clustered with CBF2, CCR2, GI, LUX and TOC1.

4.5 – Discussion

Using clustering software, genes that were expressed with a similar pattern of regulation could be identified. This was best achieved using clustering software designed to capture aspects of the expression curve rather than treating each gene as an n-dimensional data point. It was also found that due to the non-sinusoidal nature of these curves, a method that first utilised a FFT algorithm produced clusters that contained genes that were antiphase. Once SplineCluster was identified as an optimal method for clustering this data, the different sets were clustered independently.

Running the software on both individual samples and the average for each gene, it was found that repeats for each gene tended to be more similar than different genes were. This meant that most genes were separated if individual repeats were clustered. To better examine how genes compared to each other, the average of each gene was used instead as input to the clustering software. Initial analysis of cluster results showed that genes with known co-expression were clustered together in many of the appropriate conditions. Testing these expected results from the table supports the motivation for the experiment. However expanding this crude search to identify additional sets of genes that are co expressed is difficult, especially if they are not co-expressed in every condition.

To help sort out this complication, a R software package was created which would take a table of cluster identities and produce images that compared the clusters produced under multiple conditions. Initially this package had limited usability, with information being lost every time an extra condition was added. However, though adaptation of the raw code the package is now able to produce figures that keep 100% of cluster information when using up to four conditions. It still also provides information on genes that cluster together in every condition when using more conditions, but a lot of intermediary information is lost. Simple adaption could potentially extend the range of conditions that can

be visualised fairly easily, however each additional layer further complicates interpretation of the figure.

Running this software on the clustered luciferase data showed that clustering LD and LL conditions separately produced very different cluster sets. However, when clustering the full LDLL time series, elements identified in causing cluster separation in either light condition were conserved and combined to produce a larger number of clusters that fully represented gene co-expression in both a driven and free-running cycle. Examination of the figures also produced multiple interesting and novel characteristics of specific genes. These characteristics not only included identifying genes that co-cluster in each condition, but those that no longer cluster under a specific condition set. The linkage of SPA3 and ELF3 suggest that there may be a common pathway between light input (through SPA3) and the evening complex (through ELF3).

The software produced here is not limited to luciferase or spline clustered data; it was structured in a way to accept a table of any form. The only assumption is that the first column identifies the components (be it genes, metabolites, people), and that they have a value in every subsequent column. The requirement for a value in every other column could be compensated for by using a null identifier (0), which would produce a white mark instead of a coloured one. Subsequent columns then provide information on how genes compare to each other. This could be how they cluster as used within this chapter, or it could be a discrete variable (up/down regulated, present/absent, colour) provided it was inputted as a numeric. Again, this need for numerical information could easily be accounted for within the code to automate this function. As such this software provides a very useful tool for analysing large datasets that can be subdivided into groups based on a range of different information. For example, the circadian clock is known to be entrained by both light and temperature. By setting these cues out of phase with each other, genes are forced to respond either to light, temperature, or a hybrid of the two. This clustering method could identify whether genes are clustered together because of light or temperature input. By performing an experiment with light cues, an

experiment with temperature cues, and an experiment with competing light and temperature cues, 3 sets of data will be produced. By then clustering these results and applying consensus clustering, genes can be identified that are more strongly controlled by temperature/light cues.

Chapter 5 – Variational Bayesian State-Space Models Network Inference can Produce Oscillating Probabilistic Models

5.1 – Introduction

As shown previously (Chapter 4), the circadian network appears to change a lot across the temperature range, with very few components being classed as similarly expressed in every condition set. However, identifying groups of genes that appear to be co-expressed at one temperature and not another does not explain how the network topology has changed, only that it has. As such, an investigation into the topology of these genes and how they fitted together was performed. This was done using an existing network inference tool, VBSSM.

Variational Bayesian State-Space Models (VBSSM) was a network inference software package designed to take time series data and infer an underlying network (Beal et al. 2005). It worked by analysing the relationship between the (n+1)th time point of a time series and the nth time point of each of the other time series. It also allowed for the inclusion of unknown, or unmeasured, network components (hidden states). These components were important for fitting networks to incomplete or non-linear data sets as they provided an expanded framework for the model. This allowed components to be fitted together in a manner that described the data in a more complete way. The number of hidden states used to generate the output was under the control of the user. The appropriate number of hidden states to use was not self-evident prior to running the code. This was because the relative merit of using an additional hidden state could not be determined until it had been used within a simulation. As such, the software calculated the optimal network when using 0-20 hidden states; it then calculated the performance of each network, with a penalty associated with using a greater number of hidden states. The user was then asked how many hidden states they wished to use based on the result of these calculations. They were also asked to provide a p-value to use as a threshold to determine whether a connection was real or not. Any connection still present after the threshold was applied was then plotted using a Cytoscape add-on (Shannon 2003). VBSSM was an iterative code, which attempted to make small changes in each iteration and determine whether the new network better describes the data or not. This method required an initial state to be

determined, as well as some method of determining which changes to make in each iteration. These elements are derived from a string of random numbers. To ensure different combinations are used each run, and to allow reproducibility, a seed generates these random number strings.

Within this chapter, the same data described in Chapter 4.2 was used to run VBSSM simulations. The individual networks produced at different temperatures could then be compared in order to identify sections of the network that varied as the ambient temperature changed. Limitations of VBSSM limited the amount of data that could be processed at once. Firstly, processing a complete network for every gene required considerable computer time even for one seed. Running for multiple seeds usually caused a crash before data could be saved, and there was no native way to set the seed used, so simultaneous inferences could not be run. This is because the seed is set somewhere within the code, so running the inference twice will produce identical results, however running it with two seeds will produce two different networks. A reduced gene set also eased the comparison of the different. As such the decision was made to instead initially work with the core subset of genes modeled in the original circadian clock (Locke et al. 2006; LHY, CCA1, PRR7, PRR9, GI, TOC1).

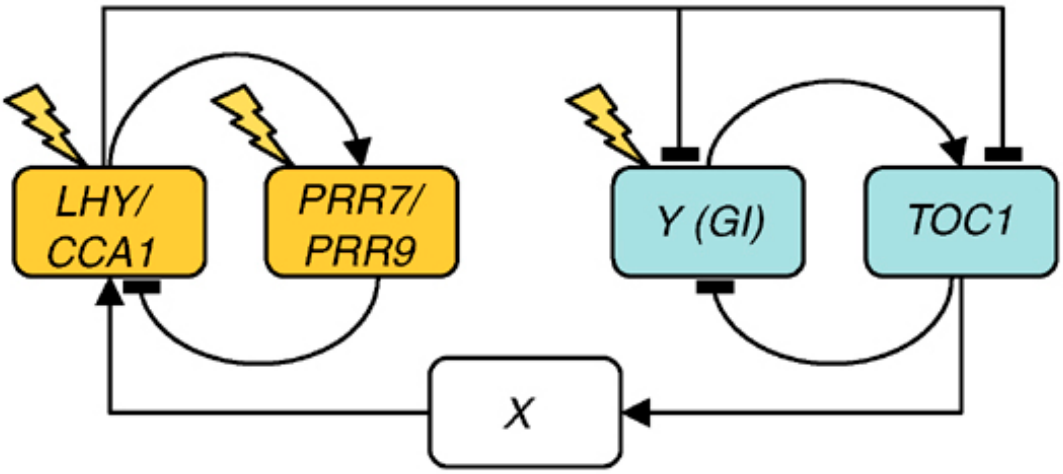
From this study, a network that matched the overall dynamics of the model was recovered at 17˚C, however this network significantly changed at other temperatures. To investigate these changes, a probabilistic Boolean model was generated and optimised. However the individual outputs of VBSSM proved too variable, so although inferred networks could be believed with enough repeats, the connection strengths used to drive simulations could not. Runs of VBSSM using different seeds generating very different output matrices caused this large variation. Despite this variation in raw values, the network of significant connections was relatively well conserved at each data set. As such, it can be concluded that the network topology undergoes significant changes in response to temperature.

5.2 – Graphical Output of Inferred Networks

To begin with, genes within the simplest circadian clock model were analysed (Locke et al. 2006 Fig 5.1 A). This was done initially at 17˚C in blue light conditions due to availability of data at the time (Fig 5.1 B). 1 seed of VBSSM, a probability threshold of 0.05 and 14 hidden states produced the network in Fig 5.1 B. The inferred network captured the core of the Locke et al. (2006) model, with LHY and CCA1 being seen to promote the PRRs, TOC1 and GI, with these in turn repressing LHY and CCA1. It also recovered the GI/TOC1 evening loop dynamics. However, there were considerably more connections than the Locke model in the inferred network. This was in part due to the separation of several of the clock components into individual elements of the network. Interestingly, though, these now separated components did not share the same regulation, nor did they always act the same on downstream components. This difference in activity of CCA1 and LHY was already known to some extend (Mizoguchi et al. 2002) and was known to be more important as temperature changes were considered (Gould et al. 2006; Salome et al. 2010). This highlighted the need to create an expanded model that simulated individual genes as single elements.

The above network inference was then repeated at 12 and 27˚C to form two more networks. When comparing the three networks, only a very small number of connections were conserved (Fig 5.2, all 3 networks are shown in Supplemental Fig 5.1). These conserved connections occurred mostly in the morning loop (promotion of PRR7 and PRR9 by LHY and repression of LHY by PRR9) although there were a few connections between the evening and morning elements (repression of LHY and CCA1 by GI and repression of GI by TOC1). Many of the connections found at 17˚C were no longer present in other temperatures and some had switched sign. Additionally, several new connections appeared in the network. Some of these changes, such as how LHY and CCA1 regulated other components, could be explained through current experimental knowledge (Salome et al. 2010), however many of the differences were not reported. It is worth noting, however, that the red auto-regulation connections might be meaningless. This was because VBSSM attempted to

A)



B)

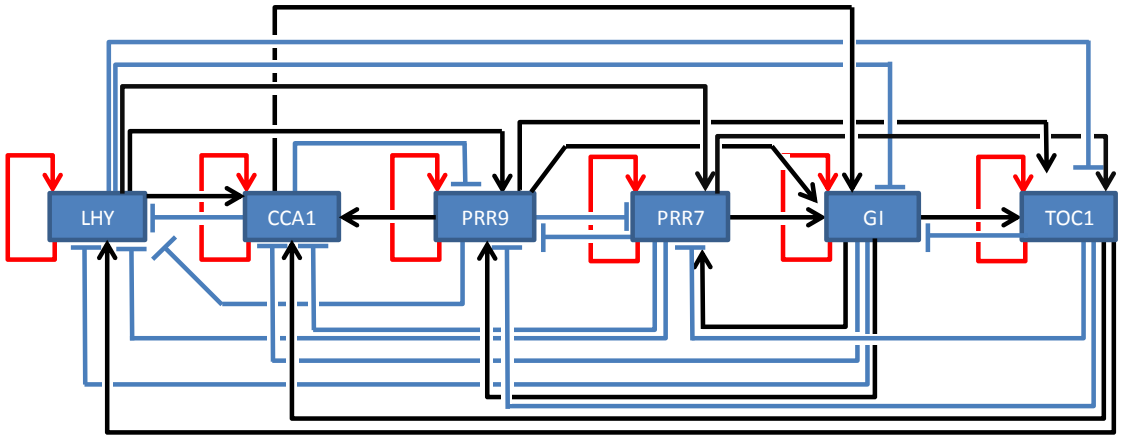


Figure 5.1 Models of the core six genes present in the circadian clock. A) Model of the circadian clock predicted in Locke et al. 2006. B) Adaptation of the Cytoscape output for VBSSM performed on 17˚C BL data. Lines are connections inferred by VBSSM coloured red for auto-regulation, blue for a repressive interaction and black for a promotion interaction.

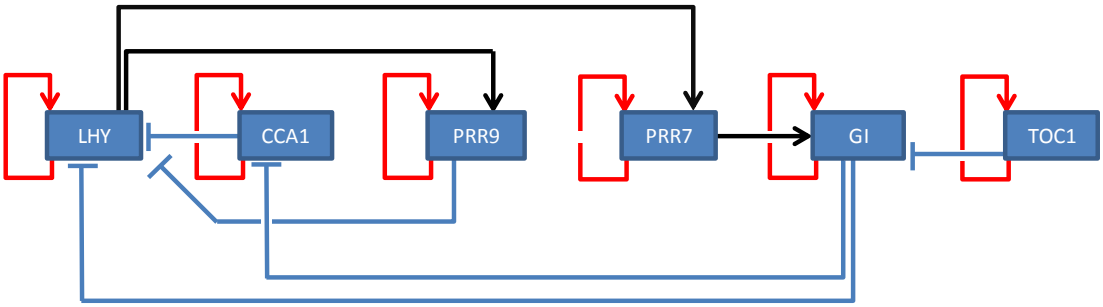


Figure 5.2 Conserved network of VBSSM inferred networks at 12, 17 and 27˚C in BL. Lines are connections inferred by VBSSM coloured red for auto-regulation, blue for a repressive interaction and black for a promotion interaction.

compare the expression of a gene at time point x+1 with the expression of each gene at time x. It then calculated how likely the expression at time x+1 was dependent on the expression at time x. Because the data being supplied is a time series, it makes sense that a gene's expression will be dependent on its previous level. However, this self-regulation was still important. It is possible for a gene to effect it's own regulation. If this repression was strong enough, the strength of this interaction would be significantly different to a gene without self interaction. VBSSM did not, however, produce a difference in these values. This suggested that either there is no self-regulation, or all the genes are under similar self-regulation. This conserved network was produced using a 5% threshold, when using a threshold of 1%, the same conserved set of connections was produced. The connections between morning components are already known to exist, further supporting this network inference. The connections from GI, however, are not published. These may, though, be caused by GI's role in stabilizing ZTL, which in turn degraded TOC1 protein (Kim et al. 2007; Mas et al. 2003). When this is combined with TOC1's role in repressing LHY/CCA1, the direct link seen in the inferred network made more sense.

With the conserved network containing so few connections compared to the individual networks, it became more likely that the entire network topology changed with temperature. To test this, genes within the network could be knocked out and temperature dependent effects could be predicted and tested. Investigating how these differences between the networks affected the overall system was difficult, however, given the number of feedback loops present. With so many ways in which different sections of the clock interacted with each other, it was hard to say what would happen if a specific gene were to be removed. In order to gain some understanding of this, a simulation for each network needed to be created. A gene mutation could then be modeled in each of the networks to see how other genes responded. These responses could then be analysed at each temperature to determine whether there was a temperature specific phenotype being produced. As such, additional information was needed to inform the networks. This information needed to include the strength of connections so that a model could be developed.

5.3 – Using Raw Output to Inform Probabilistic Boolean Networks

The raw VBSSM output consisted of a square matrix, with the list of gene names on both the column and row headings. The number in each cell then represents the strength of the interaction of gene A (row) on gene B (column). This number was referred to as a Z-score within the code. This Z-score was also used to produce a P-value, determined by how the Z-score relates to a normal distribution with mean 0 and standard deviation 1. Given the distribution of Z-scores, the values ranged from 0 to 18. This made simulation of a binary matrix difficult, so the entire matrix was rescaled to have a maximum of 1 by dividing each cell by the maximum value found within the matrix. The matrix was then further scaled by a maximum increment value (incMax) to limit the effect a gene could have on the expression of any other gene in one time step. Using this structure, a probabilistic Boolean network was then simulated (Savage et al. 2008).

Using the scaled Z-score matrix (denoted as Z), a complete interactome of all the genes analysed could be produced. This Z matrix described how the probability of gene expression (E) of each gene changed each time step. Elements of the E matrix took any value in the [0,1] range, if a value in this matrix extended past these limits, that value was replaced by the closest limit. Changes to a gene's probability occurred at times at each time step. This change was equal to the sum of all connectins onto that gene by genes that were active at the previous time step. A gene was determined as being active or inactive at each time point by considering how the probability of expression compared to a randomly generated number in the [0,1] range. This random number was regenerated for each gene independently at each time point. If this random number was less than a genes current probability of expression, the gene was declared active at that time point, if it was greater than a genes current probability of expression the gene was declared inactive. The activity state of each gene at each time point was kept in a matrix (Q), with an active gene having a value of 1 and an inactive gene having a score of 0.

Using the structures E, Q, and Z, a Boolean simulation could be run where a genes probability of expression was updated at each time point using the following equation;

$$E^i_{t+1} = E^i_t + \sum_{j=1}^{N} (Q^j_t Z^i_j) \qquad \text{Eqn. (5.1)}$$

where i was the gene being updated, j was the gene affecting i's probability of expression, t was the time point and N was the number of genes in the network.

VBSSM investigated how well the expression of a gene at time point t+1 was explained by other genes at time point t. Given equation 5.1, there will be a strong predicted interaction between a gene and itself. As such, using this central diagonal within the modeling process would unfairly represent this interaction. To fix this, the central diagonal of matrix Z, which related to self-interaction terms, were set to 0 before the matrix was scaled.

Running this equation for 1000 time steps, a time course of gene expressions was generated. This investigation was done initially at 17˚C BL with the plan to apply it to the other temperatures and light conditions once it was optimised. The gene set was also expanded to match the 11 gene Pokhilko et al. 2012 model (CCA2, COP1,ELF3, ELF4, GI, LHY, LUX, PRR7, PRR9, TOC1, ZTL), as well as including two additional reporter genes (CAB2 and CCR2).

5.3.1 – Initial Results

Using the output z-score matrix from VBSSM, which inferred using luciferase data generated at 17˚C under blue light conditions, the Boolean model was ran 20 times with initial expression probabilities (E^i_0) set randomly. The simulations generated graphs that could be grouped into four main types (Fig 5.3). Type 1 graphs (Fig 5.3 A) contained arrhythmic genes, and every gene except GI had a probability of expression equal to 0. Type 2 graphs (Fig 5.3 B)

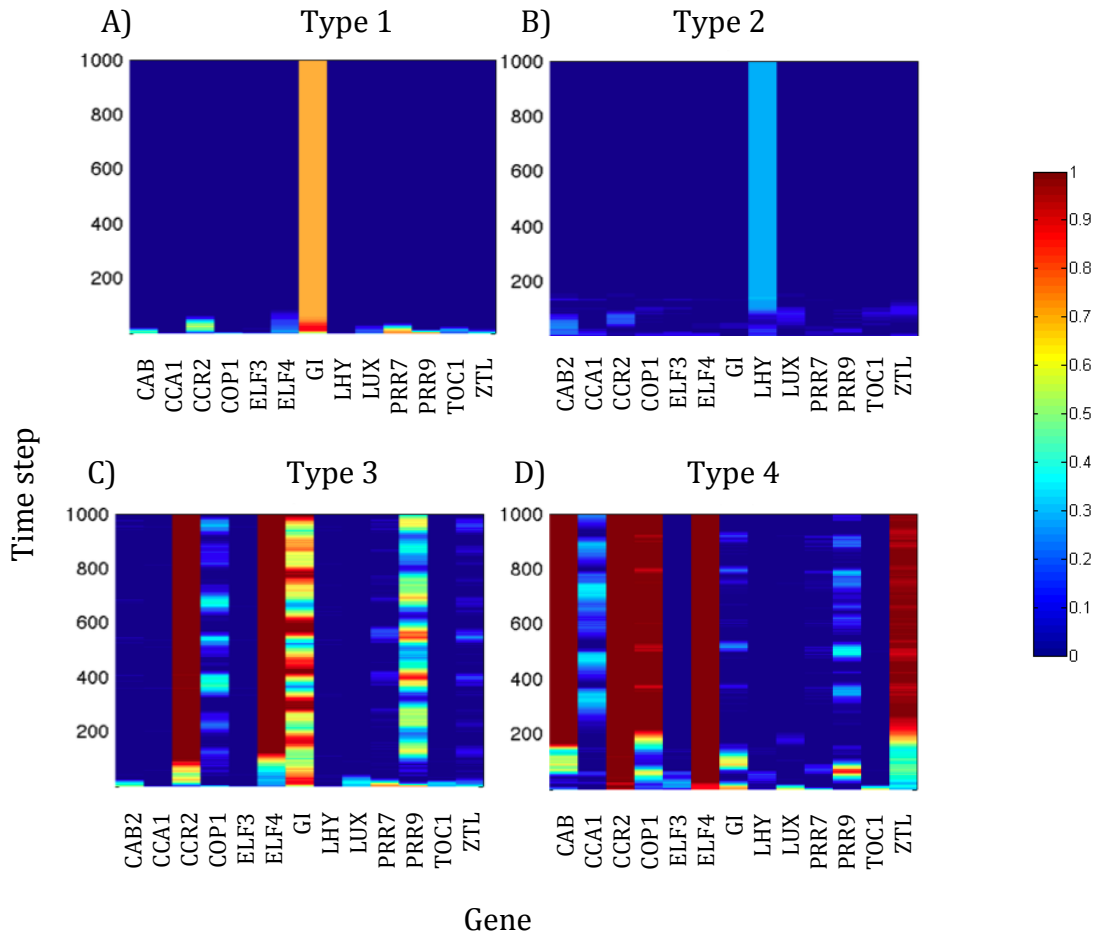


Figure 5.3 Examples of the main types of simulations recovered from VBSSM. Genes are ordered alphabetically across the x-axis and the y-axis relates to each time point simulated. When the bar is red, it means the gene's expression is near 1, if it is blue then the gene's expression is near 0. A) type 1, arrhythmic simulation with a non-0 GI expression. B) type 2, arrhythmic simulation with a non-0 LHY expression. C) type 3, rhythmic simulation with some genes demonstrating complete oscillations between the extreme values. D) type 4, rhythmic simulation where the maximum oscillation only spans half the available range.

contained arrhythmic genes, and every gene except LHY had a probability of expression equal to 0. Type 3 graph (Fig 5.3 C) contained rhythmic genes, several of which had probabilities of expression that oscillated between 0 and 1. Type 4 graphs (Fig 5.3 D) contained rhythmic genes, but the range of probabilities was rapidly reduced. Of these types of graphs, Type 3 was the most desirable, displaying features that more closely related to the original data. Types 1 and 2 were least favourable, being unable to reproduce the oscillating nature of the network.

From the sample runs it was found that 45% of runs were of type 1, 30% were type 2, 10% were a type 3 and 15% were type 4 (Fig 5.4). To investigate whether the percentage of simulations that experienced oscillations could be increased, different values to initiate the simulations were used. Rerunning the simulations, with the same starting conditions but using different seeds to determine the random number at each time step, produced results with similar proportions between the types of graphs (Supplemental Figure 5.2). This suggested that either VBSSM or the current method of simulation failed to generate a network capable of producing oscillating expressions with any starting state. Looking at the Boolean model, there were three separate ways the simulation could be adapted that might have led to more starting states producing a stable oscillation of the genes. These were; changing how the strengths of gene interactions were scaled, exploring the initial gene expression values used to seed the simulation, and varying the value of the incMax used within the simulation.

5.3.2 – Scaling the Raw Data

Up to now, the connection strength between genes had been determined purely based on the inferred Z-score, which was then linearly scaled to between 1 and 0. However, the probabilities were not linear. Z-scores were normally distributed, centred on 0 with a standard deviation of 1. As such, there was a strong argument to scale the Z-score by the p-value associated with that Z-

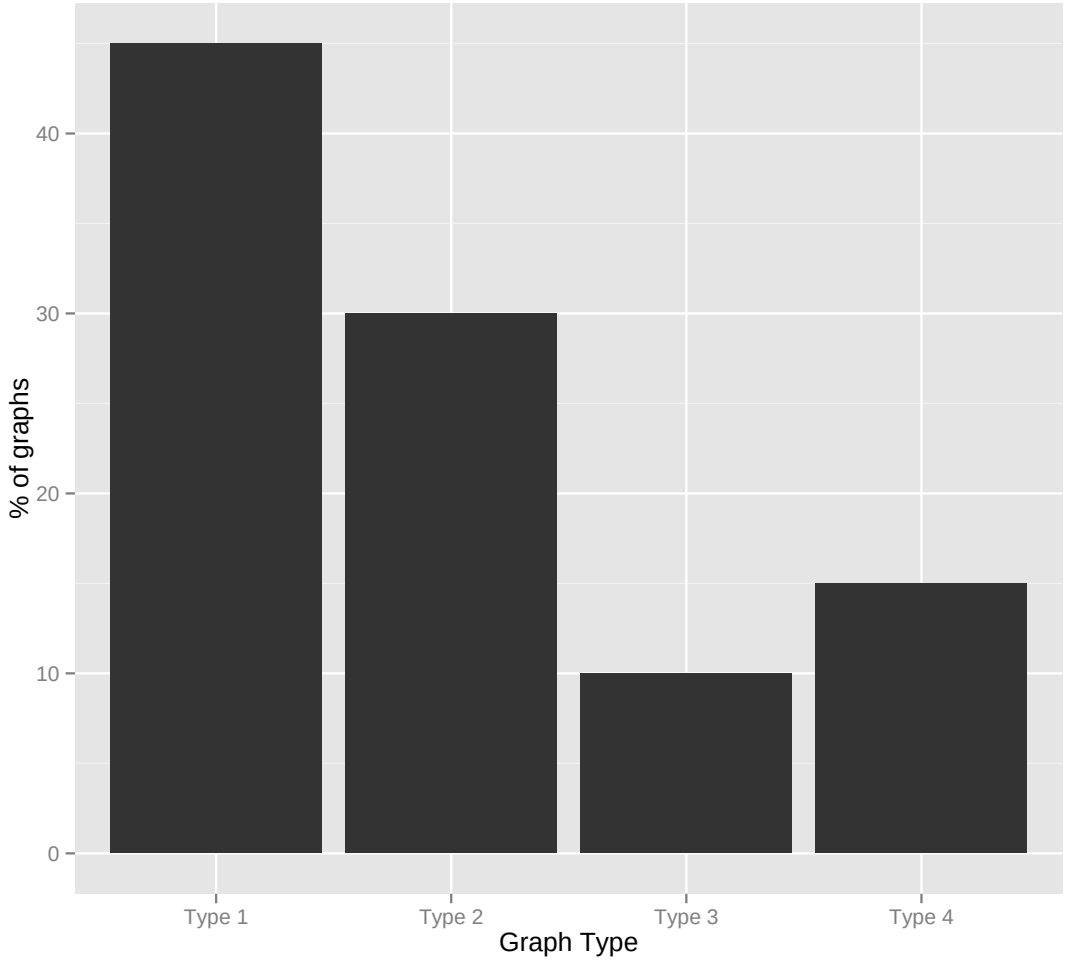


Figure 5.4 Percentage of each graph type produced by simulating the matrix produced by VBSSM. VBSSM was run using data generated at 17˚C under blue light conditions, and 14 hidden states.

score. This was accomplished by multiplying the Z-score by (1 - P-value) to create a value rescaled to better represent the significance of the connection. The use of a threshold to remove insignificant interactions from the interaction matrix was also investigated. Any time the P-value associated with the Z-score was greater than 0.05, the Z-score was set to 0. Using the same 20 simulations from above (same initial conditions and random seeds), the code was run using: 1) the P-value scaled Z-score matrix (Fig 5.5 B), 2) the Z-score matrix with threshold applied (Fig 5.5 C), and 3) the P-value scaled Z-score matrix with threshold applied (Fig 5.5 D). Fig 5.5 showed the distribution of Z-scores in these 3 cases, as well as the initial distribution as a comparison. As can be seen, there were very few visible differences when the scale was applied, however there was a large difference as soon as the threshold was applied.

Analysing the outcome of these changes, some important differences could be seen when using a threshold but little difference when scaling the connection strengths. When the scaling was applied, a previously arrhythmic simulation became moderately rhythmic (type 1 to type 4 change) and a rhythmic graph gained stronger oscillations (type 4 to type 3 change). There was also a type 1 to a type 2 switch, but this had little impact on the ability to simulate the network. When a threshold value was applied on either raw or scaled matrices, there were no type 3 graphs produced. Instead previously strong oscillations produced type 4 oscillations. When only the threshold was applied to the matrix, there were 6 type 4 graphs (30% of all simulations). When the threshold was applied to a previously scaled matrix there was 7 type 4 graphs produced (35%). These results are summarised in the bar chart in Fig 5.6.

Looking at this analysis suggested that the use of a threshold removed essential elements of the network required to maximise the range of oscillations. Additionally, whilst scaling did improve the output, it was very marginal. Running the same experiment but using different starting states in the original code could also produce this slight variation. Also, in additional simulations using different seeds, instances where the original matrix from VBSSM had produced a type 3 graph were found to produce a type 4 graph when using the

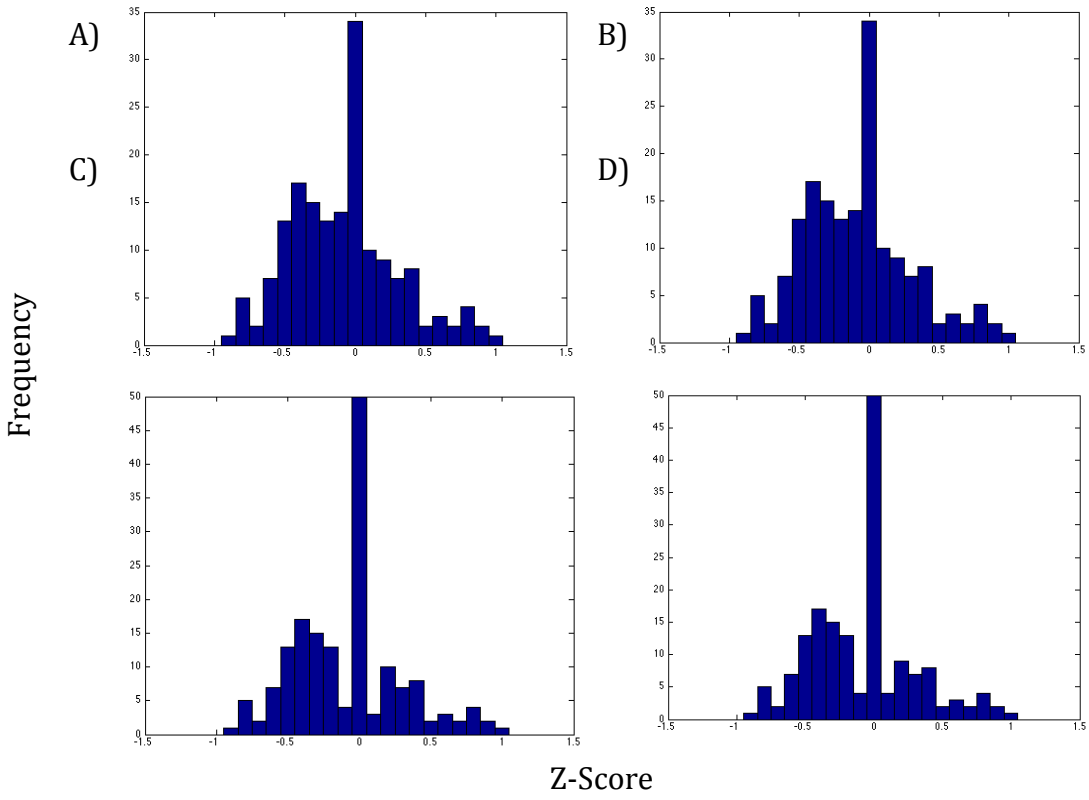


Figure 5.5 Distribution of Z-scores in the interaction matrices. The distribution of Z-scores within the matrices resulting from: A) VBBSM, B) VBSSM scaled by (1-P-value), C) VBSSM with a threshold applied, and D) VBSSM with a threshold and scaled by (1-P-value).

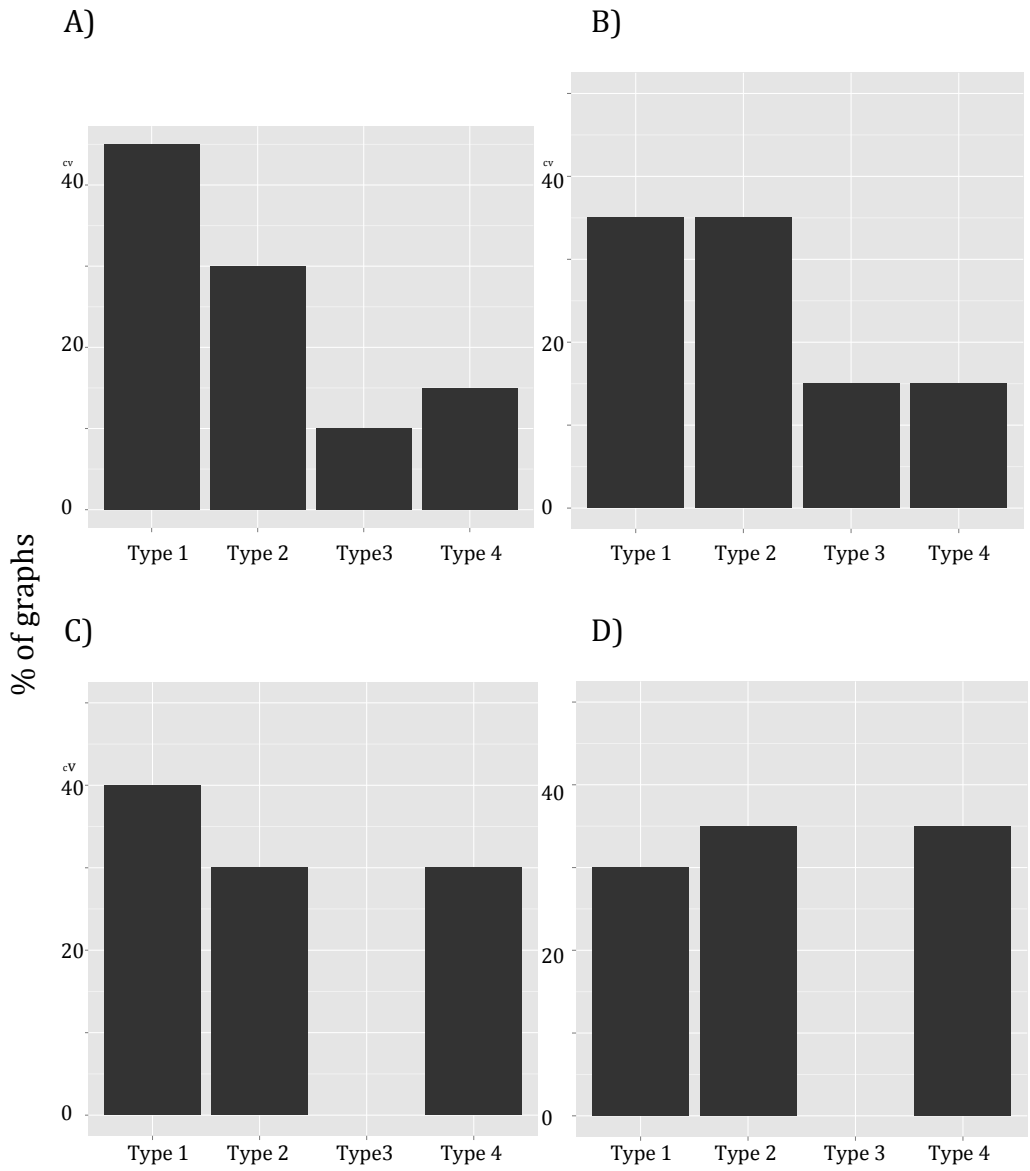


Figure 5.6 Percentage of each graph type produced by simulating the different matrices. The matrix used was created by A) VBSSM, B) scaling the VBSSM output by (1-P), C) applying a threshold to the VBSSM matrix, and D) applying a threshold to the VBSSM matrix and then scaling by (1-P). The original matrix was created by VBSSM using data generated at 17˚C under blue light conditions, and using 14 hidden states.

scaled matrix. However, the frequency of these events was also low, evident from there being no occurrences in the original experiment. From this, there appeared to be little to choose from as to how to scale the matrix, however it was obvious that applying a threshold had a negative effect to the quality of simulations.

5.3.3 – Exploring State Space for Starting Locations

To explore state space, the starting value for each of the 13 genes needed to be considered. However, even the simplest search, where every component could have one of two values, required over 8000 simulations to perform every possible combination once. Each of the 8000 simulations would then need to be run multiple times for various random seeds to get an average behaviour for each starting state. Additionally, a simple 2 state search was likely to be too simplistic, and would include only a fraction of the possible combinations. An appropriate alternative analysis would be to utilise sets of gene expressions that were relevant to the network being modeled. Given that this network was inferred from a set of data, it seamed logical that data sets from this time series should be a good place to investigate the start state space. If these starting states produced simulation outputs that infrequently contained oscillating genes, then it was unlikely the inferred network is accurately depicting the biological network.

In the dataset, genes react differently from one another when left in free running conditions compared to driven cycles. Some dampened quicker than others and some even became completely arrhythmic. Because of this, only using data from the 2 LD cycles and the first cycle in constant light was likely to provide a better understanding of how simulations responded to starting at a state generated in the luciferase screen. This was done individually using both the raw matrix output of VBSSM, and the matrix after it was scaled by (1-P). Threshold values were not applied in this investigation after having poor results in the previous tests. Each simulation was run twice, using a different random

seed. Considering the output of the individual matrices, there were no systematic differences between the resultant graphs (Supplemental Figure 5.3).

Given that there were no significant differences between the two methods, the results were pooled to provide greater sample numbers for analysis. This meant that for each of the 34 start states, there were four different simulations. By considering both a type 3 and type 4 results as successful simulations, each time point could be scored, with a successful simulation having a score of 1 and a failed simulation having a score of 0. A threshold could then be applied to the score for each initial condition (maximum score of 4) to determine which starting conditions were capable of creating consistent oscillations.

Using a threshold of 3 out of the 4 simulations, 19 of the 34 (56%) time points were found to be able to consistently produce oscillating simulations. This is a considerable improvement on the 25% of simulations originally observed (see Fig 5.6 A). Even more interestingly, there was a pattern to which start states produced successful oscillations. Oscillating start states were found to be data collected 7-17 hours after the experiment started (1st dawn), 31-45 hours (excluding 37) which relates to 7-21 hours after the 2nd dawn, and 55-65 hours which relates to 7-17 hours after entering constant light (i.e. 3rd dawn) (Fig 5.7). This result was perhaps not so surprising when the model for the 11-gene network was analysed (Pokhilko et al. 2012 Fig 1.3 C). In the model, there was a large protein complex formed from multiple genes which peak at around dusk (Evening Complex). This complex then fed back to interact with many members of the network. This complex required translational and post-translation modification data, on top of the mRNA data generated from luciferase screens, to be accurately modeled. As such the network inference for this section was likely an approximation rather than a real set of connections. Hence it was likely that simulations at this point deviated from the real data collected. Thus, when the simulation was started with real data where the EC had a prominent role in regulation, the simulation may not be able to enter oscillations reliably.

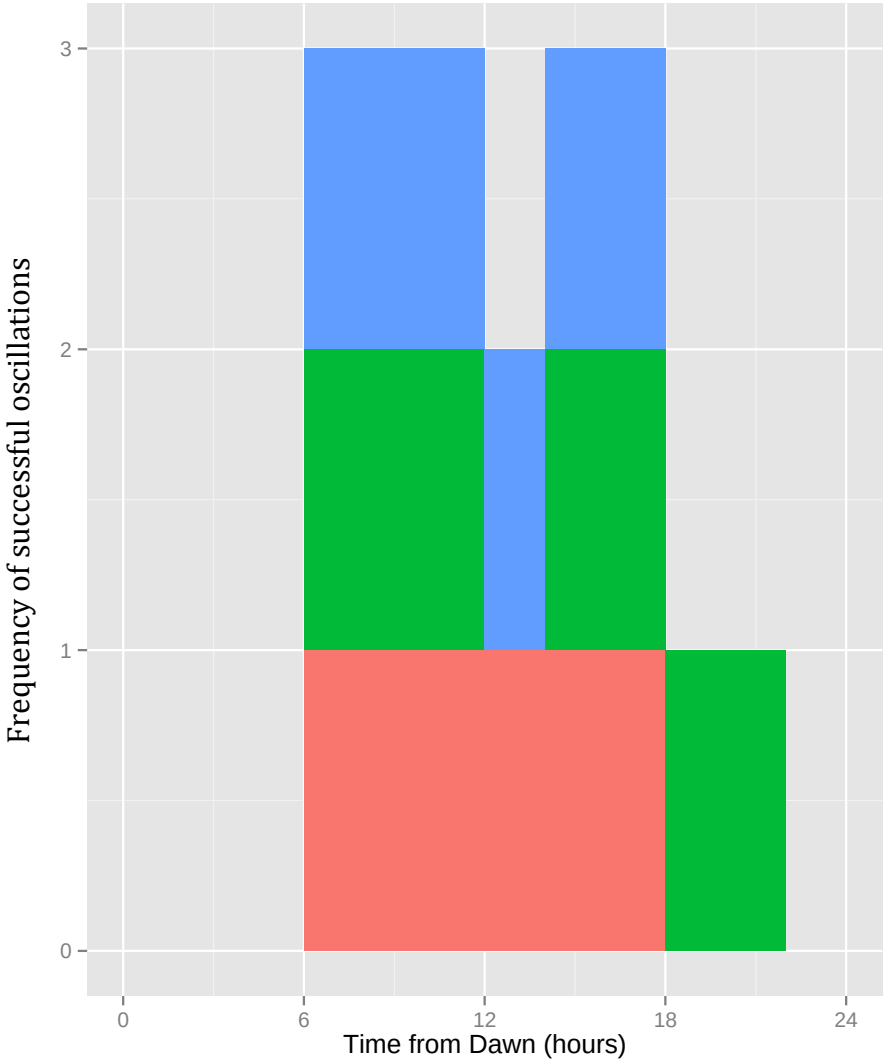


Figure 5.7 Times after dawn when simulations resulted in oscillating graphs. Pink represent the first day, green represents the second day and blue represents the third day (constant light)

From these results, it was found that using real data to start oscillations at time points where simple gene connections were thought to drive the clock was the best way to simulate inferred networks. However, four repeats may not have been sufficient, there were time points where one or two simulations created strong, type 3 oscillations but the rest were arrhythmic so that time point was called arrhythmic. Opposite to this was where there were three weak type 4 oscillations and one arrhythmic, which was called a rhythmic time point. Additionally, there were time points where one of the raw simulations and one of the scaled simulations were type 3 and the other two were type 4. This suggested there was still a large amount of variation caused by the seed used to create the random numbers that determine whether a gene was called present or absent. This led to the need to run more simulations to more accurately determine what happened under each condition set used to initiate simulations.

5.3.4 – Varying incMax

The last thing to test was how the incMax value used to linearly scale the matrix effected the simulations. Logically, the higher the value the faster the simulation oscillated (assuming it did), however there could be additional effects. Given the results above (Chapters 5.3.2 and 5.3.3), simulations were done using only the raw matrix and setting initial conditions using the data measured during the first 3 days of the experiment (34 time points). Each of the 34 starting states was simulated ten times using different random seeds to generate the random numbers. This method was applied using nine different values for the incMax: 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.

As expected, the higher the incMax, the faster oscillations occurred (Fig 5.8, D uses the original value of incMax (equivalent to Fig 5.3 C), G shows the oscillations recovered with a higher value). With incMax set to either of the two smallest values (0.005 and 0.01), the dynamics are so slow that after 1000 steps there was not sufficient information about whether a simulation had entered a stable oscillation or not. As the value increased from 0.025 to 0.4, there was a

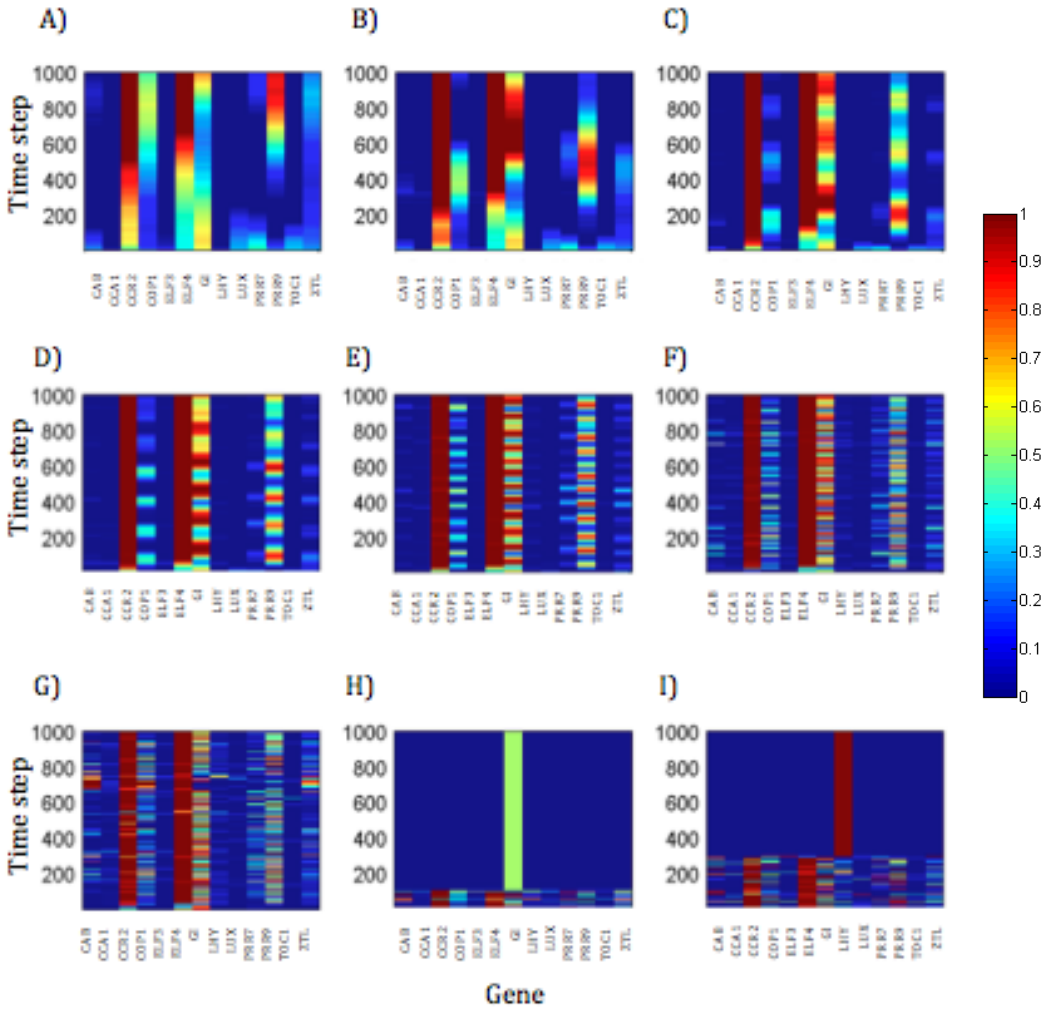


Figure 5.8 Sample result for varying incMax. Genes are ordered alphabetically across the x-axis and the y-axis relates to each time point simulated. When the line is red, it means the gene's expression is near 1, if it is blue then the gene's expression is near 0. Start data taken from time point 31.38. Simulations run with incMax equal to A) 0.005, B) 0.01, C) 0.025, D) 0.05, E) 0.1, F) 0.2, G) 0.4, H) 0.8 and I) 1.

clear increase in the speed of the oscillations (Fig 5.8 C to G). Additionally, more genes had a visible oscillation and the scale of the oscillations increased. When the value was further increased (i.e. 0.8 and 1), there became a large proportional of a new type 5 graph (Fig 5.8 H, I). These non-stable dynamics were likely the result of the random number generated at each time step coupled with the large incMax. The initial oscillations were therefore likely the result of noise in the system, persisting until most of the network had an expression probability of 0. A summary of the number of graphs of each type is shown in Fig 5.9. These figures distorted the results, however, because many of the smaller incMax simulations were oscillating in only a couple of the ten repeats in every start state. This was contrasted by higher incMax, which were more consistent in simulation results at each time point.

5.3.5 – Results

By labeling each simulation with the type of graph it produced, the effects of the incMax and start state could be examined in better detail (Fig 5.10). Oscillating dynamics did not occur for start states outside the previously described set (Chapter 5.3.2). This supported the idea that there was a limitation in how the inferred network modelled missing components, whether that was an additional gene or a protein level component (i.e. the evening complex). Simulations that were produced using an incMax equal to 0.005 and 0.01 were ignored on the basis that it was difficult to accurately assign them a type for the simulation that was produced. Looking at the remaining graphs, there was a reduction in the number of type 4 graphs (weak oscillators) as incMax increased. This was more obvious in the 3rd day, which related to data collected under constant light conditions (48-72 hours). Mirroring this was the pattern observed in the number of type 3 graphs (strong oscillations). These started relatively common in frequency, but increased in frequency as the incMax increased. However, once the incMax got above 0.8, noise dominated the simulations, and type 5 graphs became the most dominant. Interestingly, these graphs also tended to span more of the possible start states compared to types 3

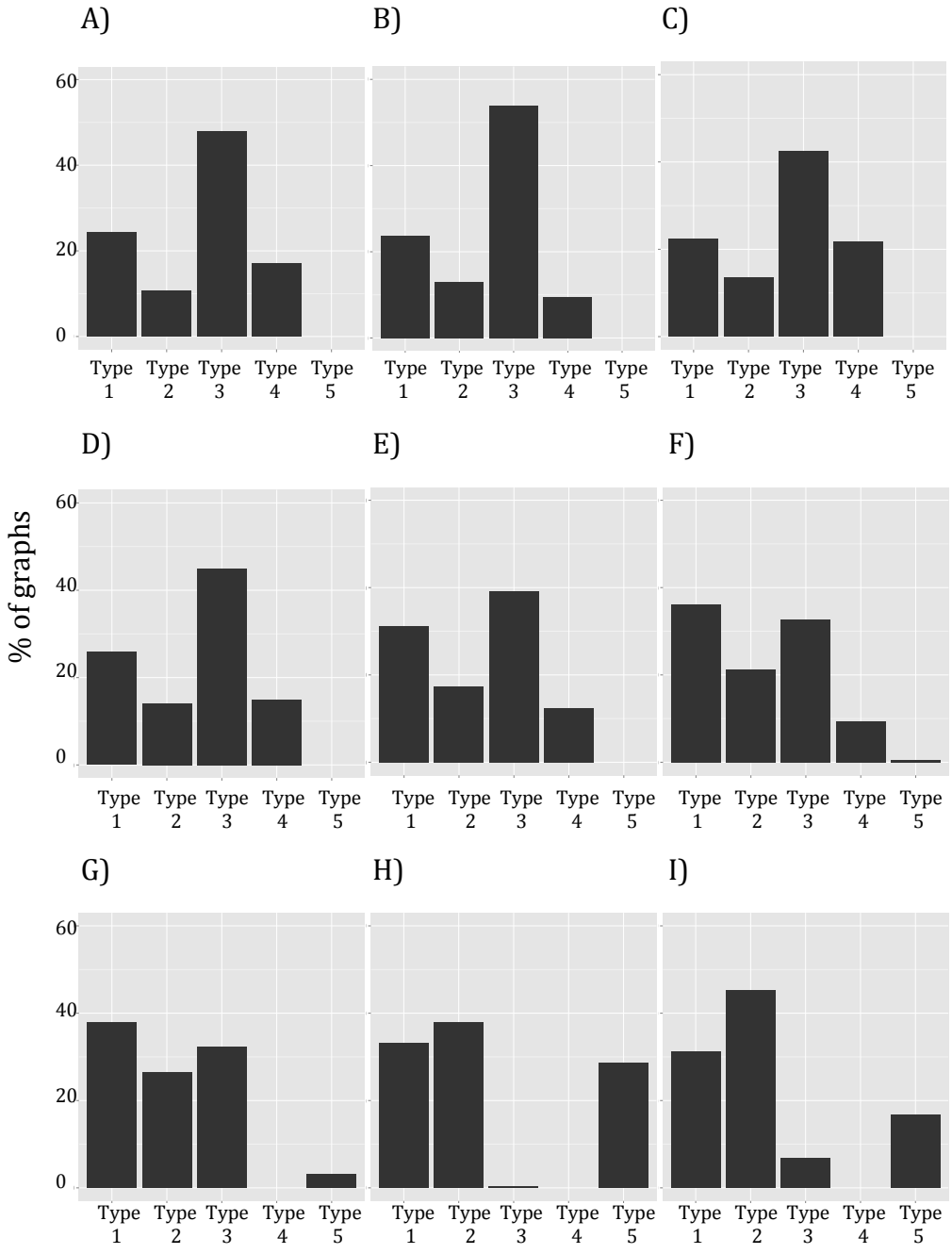


Figure 5.9 Percentage of each graph type produced by simulating the results of using various incMax values. Simulations were run with incMax equal to A) 0.005, B) 0.01, C) 0.025, D) 0.05, E) 0.1, F) 0.2, G) 0.4, H) 0.8 and I) 1.

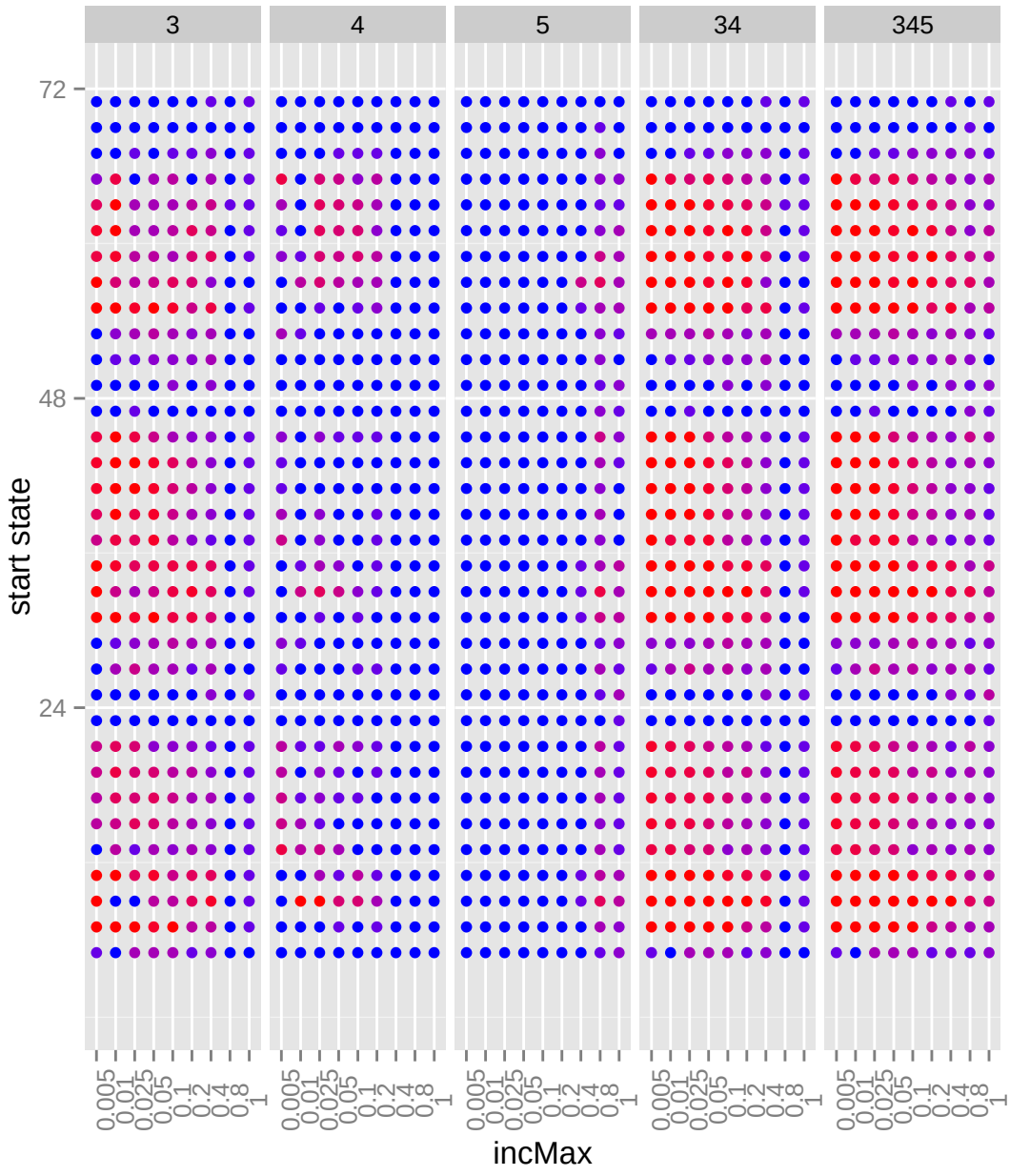


Figure 5.10 Proportion of simulations of each start state and incMax with some form of oscillation. Each of the interesting simulation types (3, 4 and 5) as well as the combination of stable oscillation (3 and 4 - 34) and summation of all interesting simulations (345) are analysed. Blue dots are simulation conditions where no simulation produces the type of graph being analysed, red dots are where all simulations produce that type of graph.

or 4. This increase in the number of simulations that reached arrhythmia may also be caused because simulations that were run with a smaller incMax had not been ran for enough to reach an arrhythmic state. From this analysis, it was decided that using an incMax of around 0.4 would produce the greatest proportion of strong oscillators, without producing the type 5 graphs. The faster oscillations seen with higher incMax also better matched the luciferase data. This data oscillated once every 12 data points, thus generating simulations with a periodicity close to this would allow better comparison between in silica and in planta expressions.

5.4 – Problem with Variation

So far, these simulations have been based on the average matrix created from four seeded runs of VBSSM. It also runs an iterative step that attempted to find a better network to fit the data. In all the previous runs of VBSSM, this value was set to 100, increasing this value would allow VBSSM to explore more possible networks in order to produce a more accurate network. However, there was considerable variation between the different repeats used to create an average. Assuming there was a specific network underlying the data that was provided to VBSSM, then the software should be able to locate it, or an approximation of it, most of the time. As such, increasing the number seeded runs used to create the average network, or how many iterations each run makes before returning an answer, should reduce this variation. This would have then resulted in a more reliable consensus network to use in modeling.

5.4.1 – Variation Between Runs

Initially, attempts were made at increasing the number of seeds used to create an average network. However, the seed used in simulations was coded deep within the structure of VBSSM and could not be easily manipulated. This meant seeds could not be run in parallel, making completion time increasingly longer. Additionally, although the seeds cannot be user defined, they were still defined in the code, so running VBSSM twice with the same data produced the same outputs. Furthermore, running with 10 seeds resulted in Matlab crashing partway through, losing all data completed thus far due to how the GUI worked. This provided little data with which to analyse how running more seeds effected the variation. As this was already being run on a high power server, a computer capable of running more seeded runs of the software could not be sourced.

Instead, VBSSM was run using different number of iterations within the code. Assuming the data has an underlying, discoverable network, then allowing the code to run for longer should have allowed it to get closer to this core network

in a more consistent manner. Again, there were limitations of what the server was able to run before it crashed mid code, however a 5 and 10 times increase in number of iterations used by VBSSM was able to be tested. Fig 5.11 showed the standard error for each individual possible connection between the 4 seeds (calculated in Matlab) when a) 100 b) 500 and c) 1000 iterations within the VBSSM code were used. As can be seen, as the number of iterations increased, so did the standard error. This suggested that the network space that VBSSM was investigating was relatively flat, meaning that there were many local minima that it could have exported as a result. Are each of these minima true aspects of a single underlying network or not? Theoretically, if each local minima (i.e. one possible network) was a true solution, then each seed should have produced a feasible model. Using each individual result to direct a Boolean model rather than using the average of four results tested this theory.

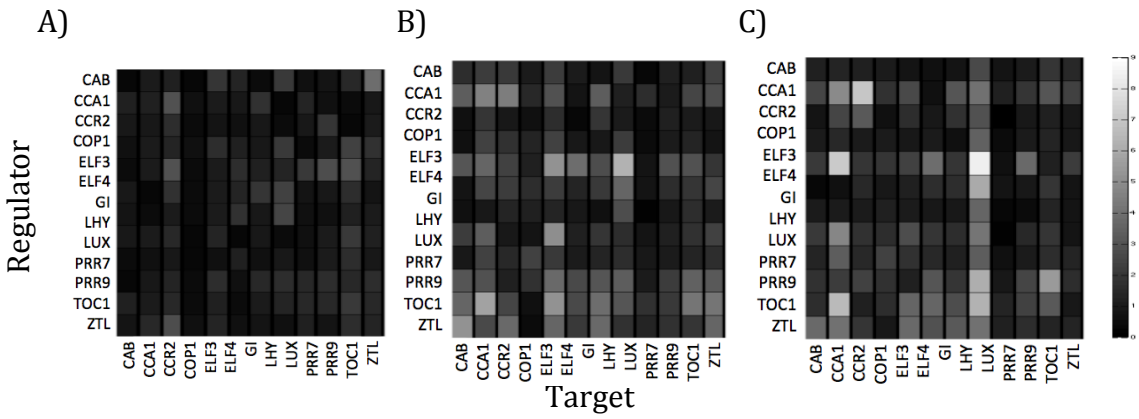


Figure 5.11 Standard error on a cell by cell basis between the four seeds from a VBSSM run. Each cell is shaded by its standard error with black meaning no error and white depicting a high error. This was done for A) 100 iterations, B) 500 iterations and C) 1000 iterations.

5.4.2 – Adaption of Modeling Method

By using the optimised method of simulating the VBSSM network discovered above (Chapter 5.3), each individual network was examined for oscillating graphs. The first thing to note was that most of the graphs were different to any of the ones seen before. This suggests that the topology of the network was significantly different between the individual outputs, compared to the topology of the averaged network. The majority of graphs produced for individual inferred networks lacked an oscillating component, however, this infrequently resulted in a type 1, 2 or 5 graph. In these simulations, a different gene was found to have a non-0, arrhythmic value. Additionally, for some graphs, several genes were found to be non-0, although were still had a constant probability of expression. This was likely due to considerable changes in the element interactions. Whilst some of these graphs had some elements that oscillated, the vast majority of genes were arrhythmic. This was especially notable in some of the time points for which the previous investigation had produced type 3 graphs in all 10 simulations. When the networks generated by individual random seeds were used (rather than the average network), only one of the networks out of four produced some form of oscillation. Indeed, none of the starting states created oscillations in all of the four individual networks, and many resulted in no oscillations at all. Additionally, only including connections in the model that were predicted as significant in three out of the four random seeds produced equally poor results in simulations. This suggested that each network was not capable of reproducing the luciferase data with any consistency, and that the successful simulations were an emergent property caused by using the average of these four seeds.

5.5 – Discussion

Here it was shown that VBSSM could produce diagrams of the networks it infers from a dataset. Using data from plants grown at 17˚C in blue light conditions, a network that contained many of the connections in the Locke et al. 2006 model was produced. VBSSM also highlighted important differences between genes commonly modeled as a single element. Networks were then produced for different datasets, and the resultant diagrams compared. From these comparisons, it could be seen that very little of the currently modeled clock was recovered in all the ambient temperatures tested. Many connections were lost and some completely changed how they modeled within the network.

The use of Boolean modeling could simulate these networks, allowing the importance of different elements to be examined. Using the average network created from four runs of VBSSM on data produced from 17˚C blue light, oscillating graphs were produced. These oscillating graphs contained only a few strongly oscillating genes, although all genes had some form oscillation. However, at an optimised value of incMax, most genes showed strong oscillations. Model analysis also demonstrated that including weak connections as well as the strong ones were required to produce oscillations with amplitudes close to the maximum. Simulations also showed that starting criteria were essential to producing oscillations. Simulations did not oscillate at times when the evening complex was heavily involved in regulation as predicted by the Pokhilko et al. (2012) model. This demonstrated that VBSSM was able to generate oscillating networks even though it was missing a significant set of data.

Variation between the random seeds, however, questioned the use of VBSSM output as input to mathematical models. A connection being consistently identified was still likely an interesting point, allowing for the construction of the network. However, using the outputted data as the basis of a model, when the value was so inconsistent, was unwise. This variation was most likely due to attempting to linearly model this oscillating system. Theoretically each of the

solutions may be a functional model that coped with the limitations in different ways. However, upon testing this, very few of the simulations produced any type of oscillation, and none of them in a consistent manner. This showed that the successful simulations originally investigated were most likely the result of an emergent property of the inferences. Due to this it seamed more sensible to attempt to use software better suited to handling oscillating data if a true measure of connection strength was to be identified.

This investigation of the luciferase data using VBSSM demonstrated that network inference tools could generate a network, which recovered significant dynamics of the existing model of the circadian clock system. It also suggested that the network topology of the circadian system was subject to significant changes as temperature changed. However, the limitation of the linear modelling that underpinned the software resulted in outputs which were heavily dependent on the seed which initiated the code.

Chapter 6 – Causal Structure Identification Infers Consistent Networks with Significant Dependence on Temperature and Light

6.1 – Introduction

Given the variable nature of connection strengths generated by VBSSM, another network inference tool was used. Causal Structure Investigation (CSI) was a successor to VBSSM (see chapter 5). The key difference, however, was how the different software packages investigated potential interactions. VBSSM assumed that gene interactions were linear in nature and used statistics based on that assumption. CSI, however, uses Gaussian statistics to determine interaction strengths. The use of Gaussian statistics allowed the interactions inferred by CSI to have a variable effect, i.e. at high levels a gene may promote the target, but at low levels it may have an inhibitory action. Given the interconnectivity predicted within the circadian clock, a gene's perceived interaction onto another gene is the combination of not only the direct interaction, but a number of indirect interactions through a different path in the network. From this, it was reasoned that a method capable of detecting interactions with a variable effect would be more likely to adjust for a complex network and would provide a better-inferred network than a method looking just for a direct interaction, as is the case for VBSSM.

CSI calculates how well a set of genes (parental set) defines another genes expression by generating an N-dimensional graph. For each time point, excluding the first, the gene being investigated has it's expression plotted on the x-axis. Each gene within the parental set then has it's expression at time point t-1 plotted against this on it's own axis (i.e. the first gene in the parental set is plotted on the y-axis, the 2nd on the z-axis etc). A surface is then fitted to this n-dimensional graph with respect to the x-axis. This surface is then scored on how well it fits the data, as well as penalised for complexity. This is done independently for each of the parental sets. Values across the parental sets are then normalized so that they sum to 1, but in an exponential manner so that small differences are made much larger. The probability that a gene is involved in the regulation of the target gene is determined by summing all of the parental set scores that it is a member of (e.g. the probability of A regulating X is the sum of AB and AC, but not BC).

CSI does not have a graphical user interface, which meant that running different data sets and changing key variables had to be done through changing the raw code within Matlab. The internal mechanics of CSI were also structured differently to those of VBSSM. VBSSM took the data and tried to produce a global network, considering the entire network as a whole, whereas CSI took each gene individually and tried to work out which other genes were involved in its regulation. This allowed for parallelisation of the code, with each gene and random seed able to be run independently. For each of these independent investigations into gene regulation, CSI created a large matrix containing all the information, formatted it, and passed it into the core mechanics of the software. Within this step, the inverse of the matrix was computed, which took a large proportion of the total run time. As such, this repeated calculation of an inverse matrix increases the total run time despite allowing parallel computing.

This software needed to be adapted in order to work with luciferase time course data (data described in Chapter 4.2, adaptation explored in Chapter 6.3.2). Additionally, variables controlling how regulators were considered, such as which genes can be promoters or the number of simultaneous connections that were to be modelled, needed to be optimised (explored in Chapter 6.3.1). Once this had been completed, results from real data and simulated data were compared (Chapter 6.4). This resulted in understanding whether it was more appropriate to extend the current circadian clock models (such as that in Pokhilko 2012) to explain the larger number of genes within this data set, or to create the entire interaction networks from scratch without taking prior knowledge into account. Using the results from these investigations, CSI was run for the data produced from each of the condition sets (light and temperature) to see how the results varied between these conditions.

After adapting CSI so that it worked with luciferase data, networks were produced with reasonable approximations to existing models. These networks were seen to change with temperature and so go some way to understanding how the circadian clock varies. However, the raw results were not suitable for simulations, making validation and further analysis difficult. An initial step to

correct for this, adding a sign to the interaction, was performed. Whilst this generated results that matched model predictions, it did not provide enough information to generate oscillating simulations when used to inform probabilistic Boolean models.

6.2 – Check of Variation

Following from the observation that variation between the random seeds used in VBSSM affected the interactions within inferred networks (Chapter 5.4), the first test with CSI was to run it multiple times and check how the different random seeds affected the output. This was done by using newly completed data for 22˚C BL and running CSI 30 times with different seeds. During this test, a 14 gene subset of the data was used. This list comprised the 11 genes named in the Pokhilko et al. 2012 model as well as CAB2, CCR2 and TIC. The 30 networks inferred by CSI were listed in random order 1000 times. These 1000 randomly ordered lists where then used to investigate whether the networks generated by CSI converged to one network. To achieve this the standard deviation between networks was defined as the sum of the standard deviations between each network connection. Standard deviations between 1000 'x-average networks' (x = 1,2,3,…30) generated from the average of the first x networks in the 1000 random ordered lists were then calculated (Fig 6.1). As can be seen, the standard deviation was very small to begin with ($3.3x10^{-7}$), nonetheless it reduced as x increased, in a typical inverse logarithmic manner. From this, it was concluded that CSI does consistently identified the same network for a given data set. The finding of a low standard deviation between 1-average networks reduced the need to run CSI multiple times per sample to create a reliable output.

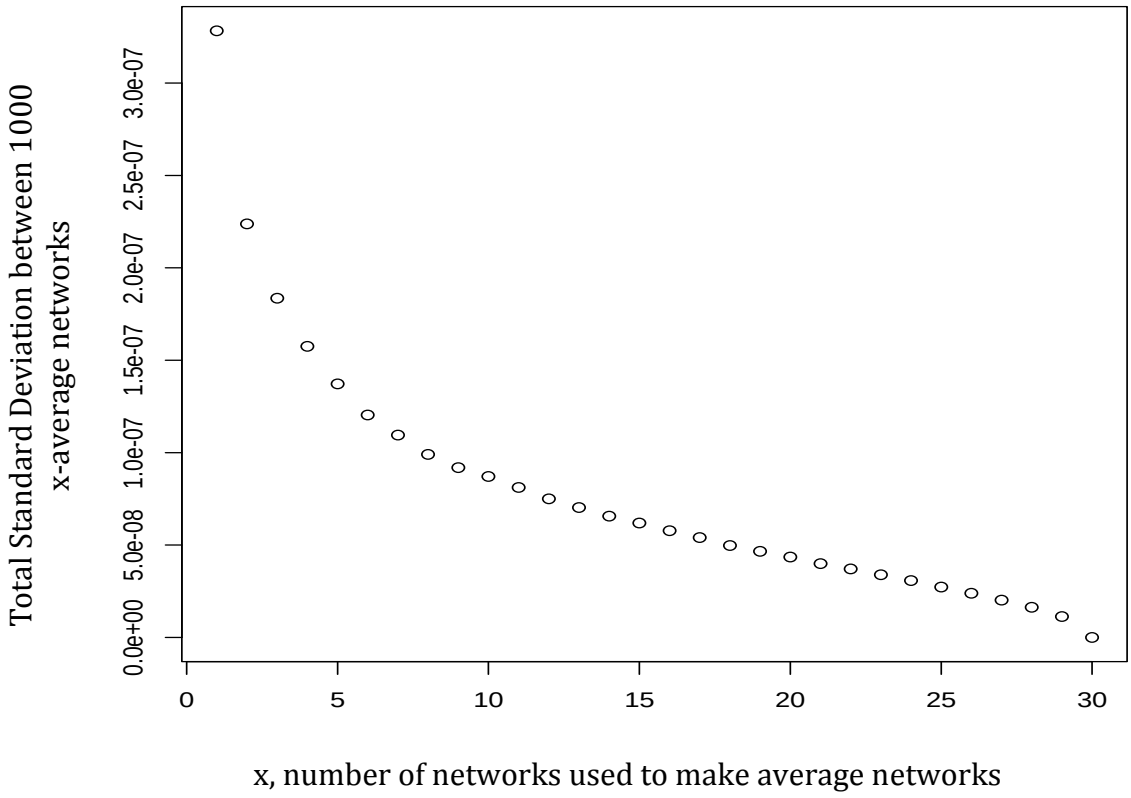


Figure 6.1 Standard deviation between networks. 30 CSI inferred networks for a 14 gene network in 22˚C blue light. The 30 networks were randomly ordered 1000 times. The x-average networks were calculated from the first x networks in each of the 1000 ordered lists. The standard deviation was then calculated between these x-average networks

6.3 – Adapting CSI for Luciferase Data

Several options within CSI needed to be selected before CSI could be run on the luciferase data. When CSI was developed, it was designed for microarray experiments, and had been optimised for sparsely connected networks. However, the data sets investigated in this work originated from luciferase screens, were more data rich, and were being used to describe a highly interconnected network.

6.3.1 – Parental Set Generation

Due to the nature of CSI, the regulators of each gene were investigated in turn. This was done by generating a set of possible parent combinations (pSet) and testing each combination from the pSet for how well it explained the genes expression. This pSet potentially consisted of every possible combination of genes in the data set. As the number of genes in the inferred network increased, the size of the pSet increased exponentially. This also led to an exponential increase in time needed for the code to complete. There were several ways the code was able to limit the size of the pSet to keep run time manageable.

The first method was to force a connection to be modeled as part of a genes regulation. By forcing a connection, it was no longer a variable, so the permutations to be included in the pSet were reduced. A good example of a connection that should be forced was self-interaction, which was assumed as a forced interaction by CSI as a default setting. This assumption originated from the simplest form of describing mRNA expression over time: mRNA levels at next time point equals the level at the previous time point plus the amount of newly transcribed mRNA minus the amount degraded within the time step. As such, within CSI, it makes sense that the level of mRNA present was strongly linked to how much mRNA was present at the last time point. Another set of connections that were considered as candidates for forced interactions were connections that existed in current models, or have been identified through

experiments. This idea of fixing a core set of connections based on a theoretical model was explored later (Chapter 6.4). The second method was to limit which genes could be included in the pSet. This could be used to only allow components known to be transcription factors to be included in the pSet. Additionally, it has already been mentioned that some of the components are expected to be outputs of the clock as supposed to being involved in regulating the clock. This method could also have been used to exclude such genes from inclusion in the pSet. However, some genes were involved in the core clock mechanism as well as having a function in the output from the clock. Limiting a gene's interaction based on current understanding of a genes role might have missed an important addition to the clock network. The third method available was the option to limit the number of simultaneous regulators within the pSet. Limiting this number reduced the number of interactions predicted as causing the expression of each gene and speeds computation. A last step in CSI before connections are given a relative strength normalised the pSet members. This normalisation scaled the probability of pSet members so that the total probability of all members was 1. Normally, one of the pSet members was significantly better at explaining a genes expression, especially after an exponential rescaling was applied. Due to how the normalisation works, this pSet member got a value of close to one and the other members had increasingly reduced values. This led to interactions within this pSet member having a relative strength of close to one, and all others having very marginal values. As such, each gene was predicted to be regulated by the number of connection present in the optimal pSet member. Thus, if the pSet was limited to 2 simultaneous connections, a maximum of two connections were likely to be predicted for each gene. To determine what the optimal maximum number of parents (cMax) should be in this investigation, a subset of genes from the data was passed through CSI with multiple cMax values. In Chapter 6.4, networks inferred from real data were compared to those from model-simulated data (Pokhilko et al. 2012, Domijan et al. 2014). As such, it was logical to use an equivalent data set from the luciferase data to investigate how varying cMax altered the inferred network. By using information about the same set genes, a direct comparison between the networks produced by CSI for the simulated and

luciferase data could be made. It also allowed for the analysis of how the cMax value affected the networks produced from each of these data sets. The simulated data contained mRNA expression levels for eight different genes within the circadian network (LHY, PRR9, PRR7, TOC1, ELF4, ELF3, LUX, GI). The luciferase data for these genes was inputted into CSI with a forced self-interaction. This was repeated independently using cMax values from 1 to 5 (Table 6.1). The results in Table 6.1 show that as cMax was increased, the connections predicted for smaller values of cMax were retained. The exception to this pattern was sometimes seen when cMax increased from 1 to 2. There was several times where the connection predicted using cMax =1 did not appear again in networks generated with higher values of cMax (e.g. PRR9 regulating GI). The pattern of components remaining as cMax increased also suggested that there was an order in how important each connection was. Information on the relative strength of connections was lost, since all connections in the pSet had a value of approximately 1 (due to the scaling process described above). There was one exception to the bimodal value (0,1) distribution, ELF4's predicted regulation, using cMax=2. Here the ELF4 regulators PRR9 and ELF3 had fractional values, which added to 1, in addition to GI having a value of 1. This suggested that if more than cMax connections were necessary to explain the network, or if two connections were equally able to explain the expression seen, the code was able to display it. There were also several examples where increasing cMax did not alter the number of predicted connections (most clearly seen for PRR9, who's output did not change after cMax was increased from 3). This was promising as it showed that CSI did not just return a complete interconnected network if cMax was not limiting. From this analysis, it was decided that using a cMax of 2 would result in obtaining the most important, reliable results within a reasonable computational time frame, which was essential consideration when the number of components was increased. As they were generally preserved as cMax increased, the connections predicted using a cMax=2 are the most reliable.

Table 6.1 Predicted regulators of genes within Pokhilko 2012 SaSSY model. Connections were inferred using luciferase data generated in 22˚C blue light

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
LHY	PRR9	PRR9	PRR9	PRR9	PRR9
		GI	GI	GI	GI
			ELF3	ELF3	ELF3
				PRR7	PRR7
					TOC1

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
PRR9	ELF3	ELF3	ELF3	ELF3	ELF3
		GI	GI	GI	GI
			TOC1	TOC1	TOC1

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
PRR7	PRR9	TOC1	TOC1	TOC1	TOC1
		ELF3	ELF3	ELF3	ELF3
			GI	GI	GI
				PRR9	PRR9
					ELF4

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
TOC1	LHY	GI	GI	GI	GI
		ELF3	ELF3	ELF3	ELF3
			ELF4	ELF4	ELF4
				PRR9	PRR9

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
ELF4	GI	GI	GI	GI	GI
		PRR9 0.7	ELF3	ELF3	ELF3
		ELF3 0.3	TOC1	TOC1	TOC1
				PRR7	PRR7

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
ELF3	PRR9	PRR9	PRR9	PRR9	PRR9
		GI	GI	GI	GI
			TOC1	TOC1	TOC1
				PRR7	PRR7

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
LUX	GI	GI	GI	GI	GI
		PRR9	PRR9	PRR7	PRR7
			ELF4	ELF4	ELF4
				ELF3	ELF3
					TOC1

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
GI	PRR9	PRR7	PRR7	PRR7	PRR7
		ELF4	ELF4	ELF4	ELF4
			ELF3	ELF3	ELF3
					LHY
					TOC1

6.3.2 – Use of Luciferase Expression Data

CSI was designed around using microarray data for inferring a network. Whilst the nature of the data is similar, with both microarrays and luciferase screens providing a means of determining the mRNA expression levels, there was one key difference: In a microarray, the data gathered for each gene within a sample was fundamentally related i.e. the same genetic material was used to measure the expression of each gene. However in a luciferase screen, each gene was measured using a different construct inserted into a different plant. Whilst technically the plants should behave the same, biological variation may have made one plant run slightly fast and another slightly slow. This was potentially an issue because CSI worked by assuming that the genes in each replicate were intrinsically linked, with the data coming from the same plant. Thus the luciferase dataset used as input to CSI may drastically alter the output, depending on which gene replicates were combined to produce a single 'plant'. The question then became: if the repeats were mixed up to produce different 'plants', would the results have been the same?

Previously, the luciferase repeats were used as the expression time series for each gene. In the order they were placed within the luciferase experiment. In this investigation, for each gene, time series from the set of repeats were chosen randomly, with replacement, 8 times. These randomised luciferase repeats were used to construct the CSI input dataset. This protocol was repeated 30 times. Everything else was kept the same between the runs, including the random seed within the core CSI software. The output of this was analysed in the same way as the seed used to initiate CSI repeats were analysed in Chapter 6.2 (Fig 6.2). Whilst the same rapid decrease in standard deviation could be seen compared to Figure 6.1, it started at a far higher value (25 compared to 3.3×10^{-7}). This showed that the variation between how the data was combined before it was fed into CSI was far more important than the seed used to initiate CSI.

The effect of using different luciferase randomized averages was more obvious when considering the inferred connections of each component. Fig 6.3 showed

the distribution of inputs into CAB2 from other genes in the network (and itself, leftmost point on each graph). Although the code was limited to only 2 non-self interactions, for 1-average networks 7 different genes were reported to input CAB2 in at least one of the networks. As more networks were used to create the x-average network, many of these 7 CAB2 inputs rapidly reduced in strength. Variation around the mean point also decreased, showing that as x increased, the x-average networks converged onto a single network. Using the average of 30 randomised luciferase sets not only removed the bias of using a single arbitrary combination, but also provided information about how important each connection was in a manner that was not bimodal or limited by the cMax. It was known that the network was more complex than two transcription factors affecting each gene, but as previously mentioned increasing the number of simultaneous connections would have greatly increased computing time. Using this method, more than two genes were recovered as potential regulators, and additionally they could be ordered in terms of importance (Fig 6.4). This is an ability lost when CSI was run with a higher cMax. Regulator graphs for each gene were shown in Supplemental Figure 6.1.

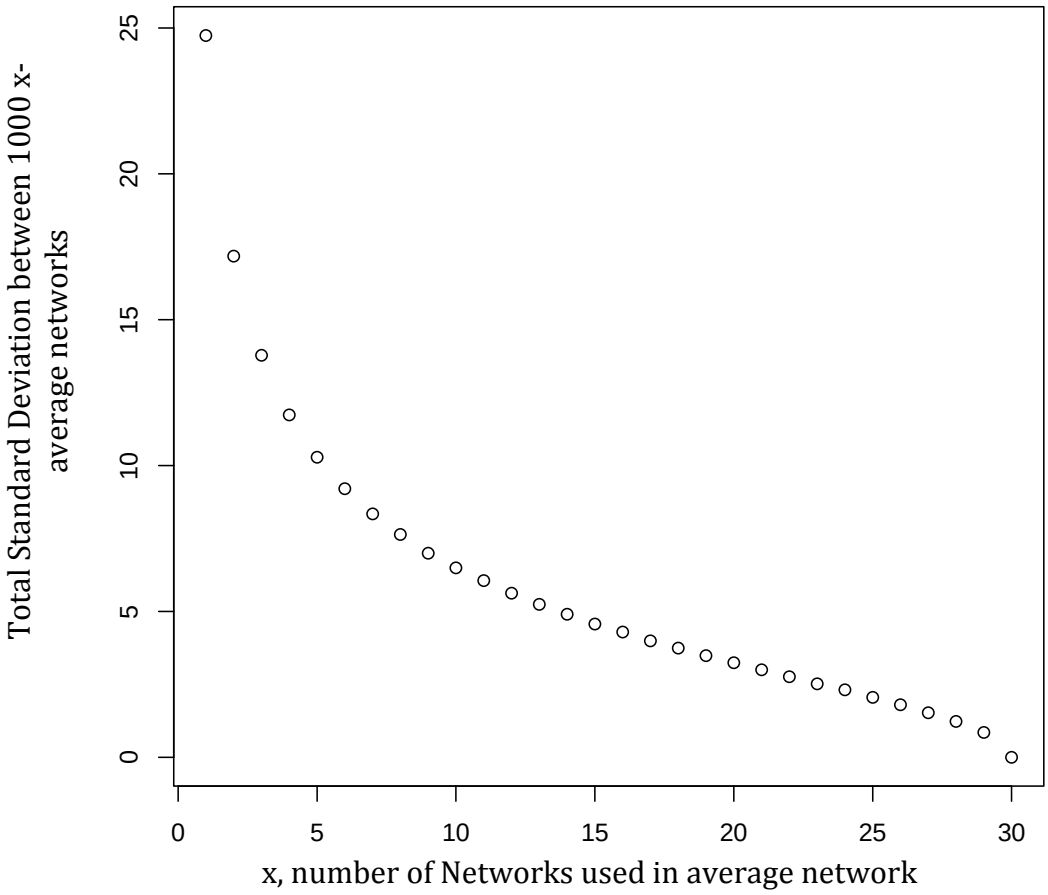


Figure 6.2 Standard deviation between networks. 30 CSI inferred networks for a 14 gene network in 22˚C blue light. The 30 networks were randomly ordered 1000 times.. The x-average networks were calculated from the first x networks in each of the 1000 ordered lists. The standard deviation was then calculated between these x-average networks

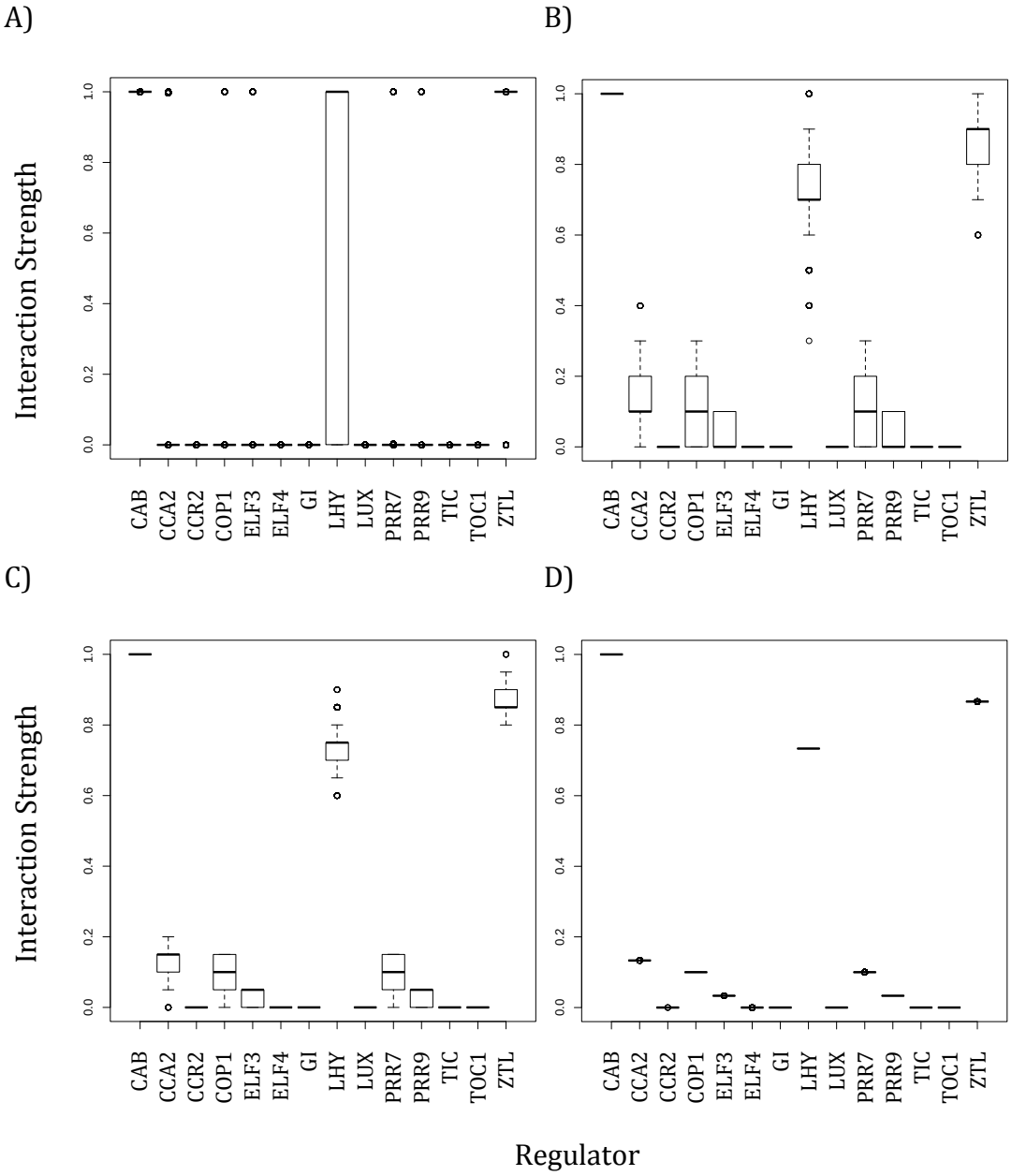


Figure 6.3 Distribution of interaction strengths outputted by CSI using 30 randomised luciferase sets. x-average networks are created by averaging A) 1, B) 10, C) 20 and D) 30 networks, repeated 1000 times. Graphs show the distribution of inputs into CAB2 from other genes in the network (and itself, leftmost point on each graph). Boxplots, dots are outliers, thick black line is median of distribution, and boxes cover the middle 50 percentiles of the distribution.

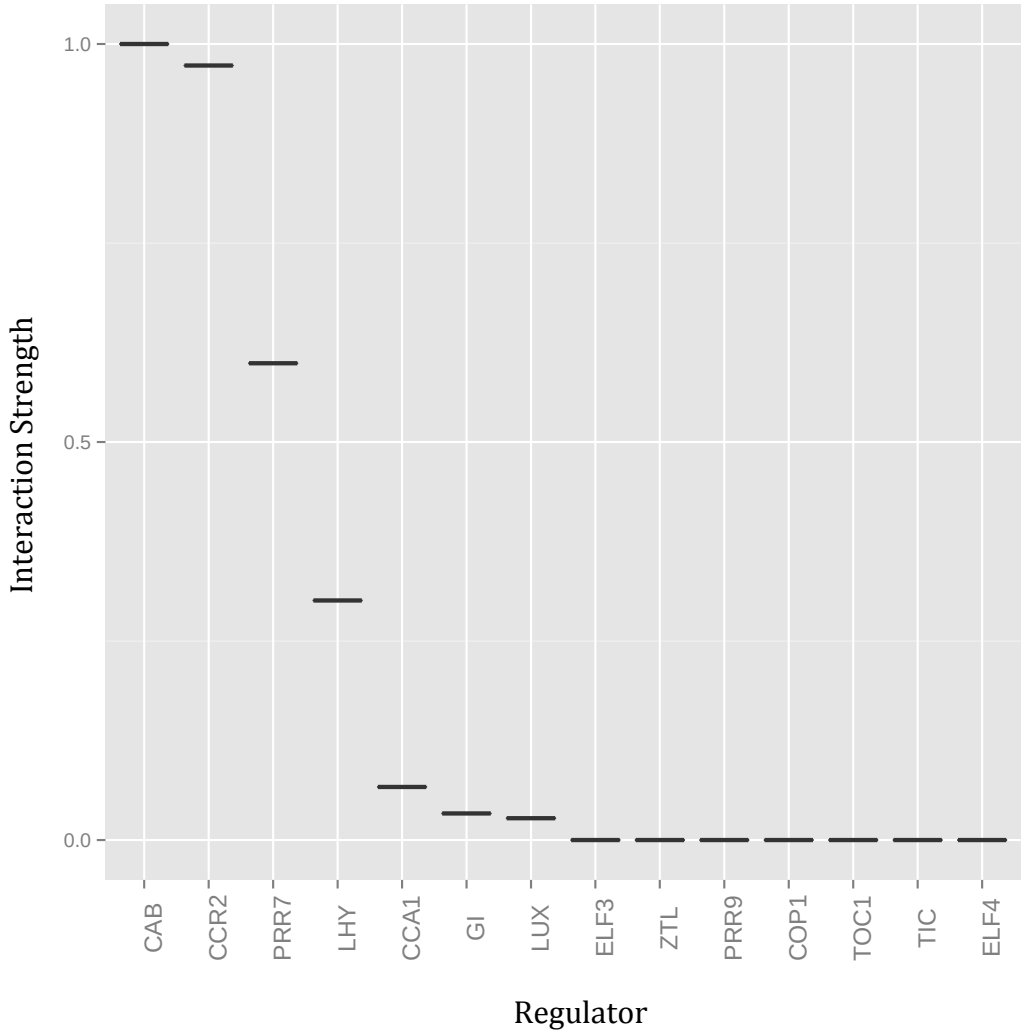


Figure 6.4 Distribution of interaction strengths outputted by CSI using 30 randomised luciferase sets, ordered by interaction strength. Graphs show the distribution of inputs into CAB2 from other genes in the network (and itself, leftmost point).

6.4 – Applying CSI to Simulated Data

If the core network of the circadian clock were conserved across the conditions, then this network would most likely be pulled out by CSI in each of the conditions. Not forcing these core network connections may result in few differences between the different networks and therefore produce very little information about condition dependent network changes. As such, it would be more informative to force the core connections. Assuming that the current model of the circadian clock (Pokhilko et al. 2012) was the foundation of each of the networks, then the results of using CSI on the luciferase data should match very closely the results of using CSI on Pokhilko et al. 2012 simulated data.

6.4.1 – Luciferase Data vs. Pokhilko 2012 Simulation

To test this idea, data from a simulated model (Pokhilko et al. 2012) was passed through CSI (Table 6.2), only simulated profiles of components previously analysed in Chapter 6.3.1 (Table 6.1) were used. Network inference of the simulated data recovered a high number of the interactions found in the Pokhilko et al. 2012 model (Figure 1.3 C). Limitations of the model meant that some genes were missing from the network, as they were not simulated at mRNA level. Additionally some interactions are incorrectly modeled by CSI since none of the information about protein or protein complexes from the simulation was supplied to CSI.

Comparing table 6.1 with 6.2, there were many differences, apparent by lack of highlighting in Table 6.2. A major difference between the data in tables 6.1 and 6.2 was how they acted at high values of cMax. In the luciferase networks, the majority of gene regulators increased with cMax. However, in the networks inferred from simulated data, all genes have reached their maximum number of regulators by cMax =3. Additionally, although many of the connections predicted by simulated data were also predicted by the luciferase data, these

Table 6.2 Predicted regulators of genes within Pokhilko 2012 SaSSY model. Connections were inferred using data simulated under the same conditions as the experiment. Cells highlighted in green show interactions recovered by CSI networks inferred using luciferase data (Table 6.1) as well as those inferred using simulated data.

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
LHY	TOC1	PRR7	PRR7	PRR7	PRR7
		ELF3	ELF3	ELF3	ELF3

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
PRR9	LHY 0.8	LHY	LHY	LHY	LHY
	GI 0.2				

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
PRR7	LHY	LHY	LHY	LHY	LHY
		ELF3	ELF3	ELF3	ELF3

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
TOC1	ELF4 0.5	GI	GI	GI	GI
	LUX 0.5	PRR9	PRR9	PRR9	PRR9

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
ELF4	GI	GI	GI	GI	GI
		PRR9 0.5	PRR9 0.3	PRR9 0.3	PRR9 0.3
		ELF3 0.5	ELF3 0.7	ELF3 0.7	ELF3 0.7

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
ELF3	LHY	LHY	LHY	LHY	LHY
		PRR7	PRR7	PRR7	PRR7

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
LUX	GI	GI	GI	GI	GI
		PRR9 0.5	PRR9 0.3	PRR9 0.3	PRR9 0.3
		ELF3 0.5	ELF3 0.7	ELF3 0.7	ELF3 0.7
			ELF4 0.2	ELF4 0.2	ELF4 0.2

Gene	cMax=1	cMax=2	cMax=3	cMax=4	cMax=5
GI	ELF4 0.3	ELF4 0.5	ELF4 0.5	ELF4 0.5	ELF4 0.5
	ELF3 0.3	ELF3 0.16	ELF3 0.16	ELF3 0.16	ELF3 0.16
	LUX 0.3	LUX 0.5	LUX 0.5	LUX 0.5	LUX 0.5
		PRR9 0.85	PRR9 0.85	PRR9 0.85	PRR9 0.85

usually only featured when CSI was run with high values of cMax on the luciferase data. This potentially suggested that the biological network, which controls gene expression, was significantly different to the model in terms of the most important regulators in a topological study. Additionally, the simulations lacked mRNA information for several key components within the clock. For example CCA1 was completely missing from the Pokhilko et al. (2012) model as it was assumed to be the same as LHY for modeling purposes. Additionally, although ZTL was modeled in the Pokhilko et al. (2012) clock, it was only done so at a protein level. From this analysis it was seen that forcing mRNA network interactions, within CSI, for the analysis of luciferase data, informed by the Pokhilko et al. (2012) model, which was itself informed by biological information at multiple levels, is unwise and likely to have required even more assumptions of the biological system. Thus, it made more sense to use CSI to fit networks to each condition set independently, without a set of prior assumptions. Had this have resulted in a single core network in all conditions, than the core interaction network would have been fixed in CSI and the software rerun.

This investigation did provide some very interesting information about how CSI coped with the evening complex protein complex. For instance, GI was modeled in Pokhilko et al. (2012) as being repressed by the evening complex. The evening complex is made up of ELF3, ELF4 and LUX. After running CSI on the Pokhilko et al. 2012 simulated data, all three components of the evening complex were predicted as having a role in regulating GI. However, this was not done in the binary format normally returned by CSI, these partial connections were maintained even when cMax was increased to a point that would allow all of the connections to have a value of 1. This shows that a pSet member containing all three components as well as the additional interactions did not provide an increased explanation to a genes expression however any of them on their own was able to describe the connection with similar effectiveness.

6.5 – Recreating Underlying Networks

Above, it was shown that networks created from luciferase data did not fully match the networks produced using simulated data. However, the results of both networks produced elements that resembled the central network of the circadian clock. Additionally, CSI and other network inference software (such as DBN's (Werhli et al. 2006)) have been able to produce approximations of the simple 6-gene model (Penfold & Wild 2011; Pokhilko et al. 2010) using microarray data. A good starting step to determining whether the luciferase data set can inform meaningful networks was to determine if a network similar to those already published could be recovered. In Penfold & Wild (2011), network inference was done by running the software on the 6 genes in the Pokhilko et al. (2010) network, as well as adding ELF4, which had been used in the previous modeling study that generated the microarrays (Zou et al. 2009). CCA1 and LHY were then combined by taking the union of their inferred connections (maximum strength of the connections being combined), as were PRR7 and PRR9. This was done to better represent the elements modeled in the Pokhilko et al. (2010) (Fig 6.5 A). Once the results from luciferase data at 22˚C blue light had been compared to the relevant networks produced in the paper (Fig 6.5), the process was repeated for the other conditions to see how temperature and light affected the results. Since Penfold & Wild (2011), a new model for the circadian clock had been generated, with more components included within the network (Pokhilko et al. 2012). As such, the inference in this study was then expanded to see how the newer 11-gene network (Pokhilko et al. 2012) compared to inference of the former 6-gene network. Additional clock component, TIC (Hall 2003), as well a couple of reporter genes, CAB2 and CCR2 (Millar & Kay 1991; Carpenter et al. 1994), were also added to the data analysed, to see how their presence changed the network topology compared to existing models.

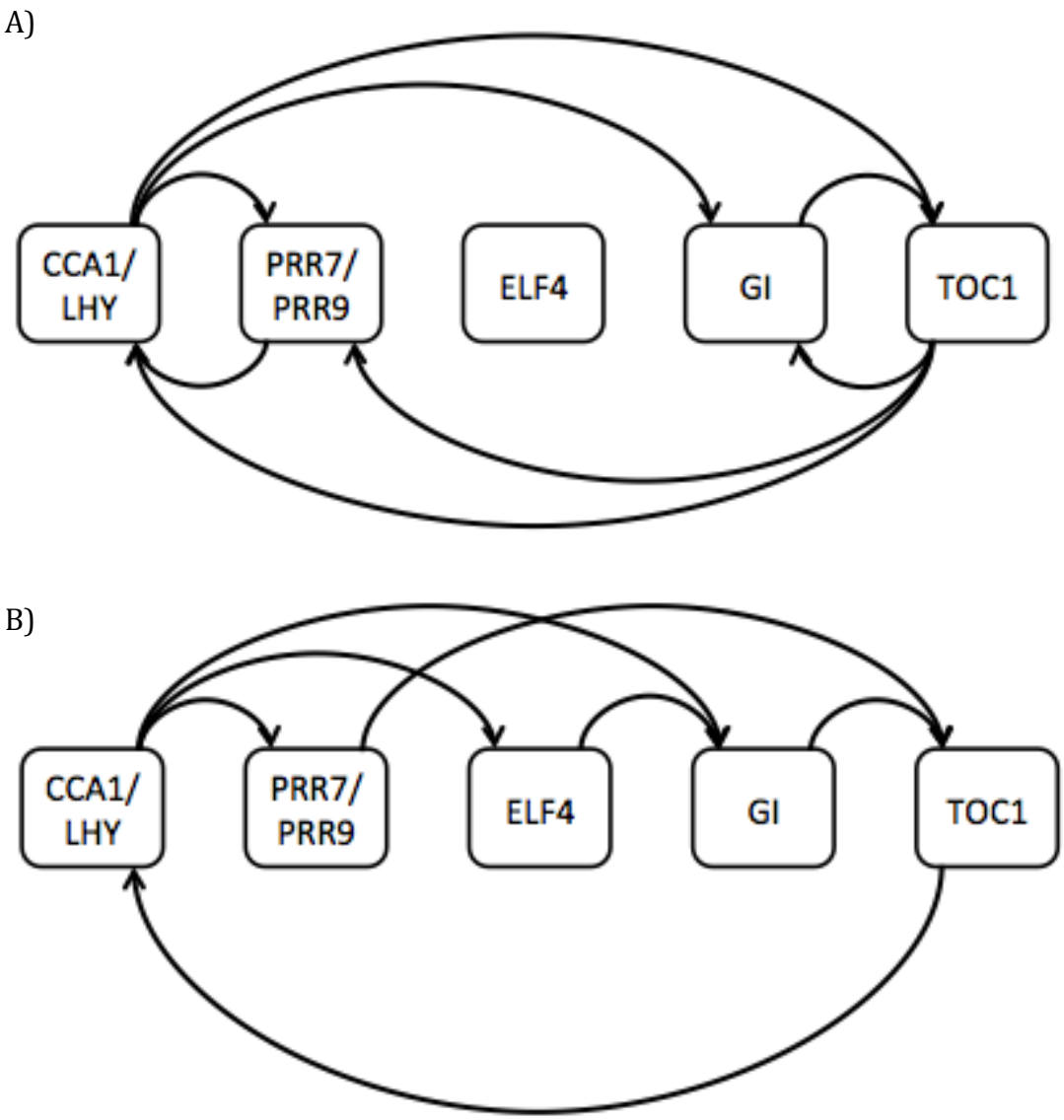


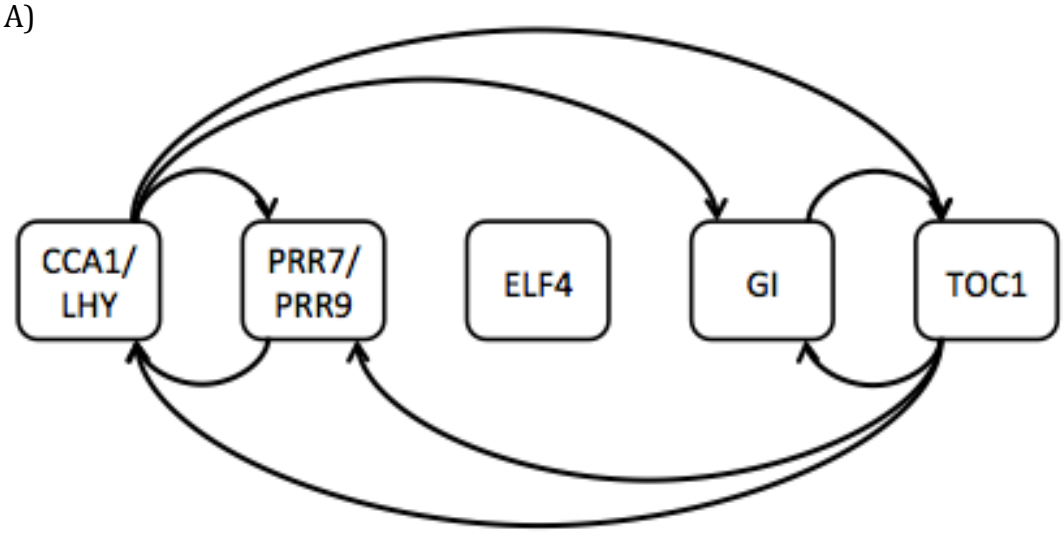
Figure 6.5 Results of CSI inference on microarray data (Penfold & Wild 2011) compared with the Pokhilko et al. 2010 model. A) A modified image of the Pokhilko et al. 2010 circadian clock model. B) The results of using CSI on microarray data generated at 22˚C in white light (adapted from Penfold & Wild 2011).

6.5.1 – 7 Component 2006 Network

To compare the ability of luciferase data to produce a network equivalent to a predicted model (such as Pokhilko et al. 2010), or a network constructed by microarray data, CSI was run in a similar manner to Penfold & Wild 2011. This was done by using the data collected at 22˚C in blue light, with a cMax of 2, no forced interactions other than self-interaction, and no restrictions on which genes were able to be in the pSet. CSI was then run 30 times using different combinations of repeats to form individual 'plants' (as explained in section 6.3.2). These were then averaged to produce one network and the union of the outputs for CCA1 and LHY was calculated as well as that for PRR7 and PRR9. By applying a threshold to the matrix containing the connection strengths of each possible connection, the network can be reduced to include only the strongest connections. This threshold was set so that the top 10 connections were used to create the network (Fig 6.6). This required applying a threshold of around 0.45 to the matrix. When this method was applied to the other conditions, the threshold needed to generate the top 10 connections changed. Alternatively, a set threshold could have been applied to each condition, which would have produced a different number of connections for each network.

Instead of creating different thresholds to produce a specific number of connections, or applying an arbitrary threshold to all the networks based on the strength of the tenth component in one of the conditions, exploring how the connection strengths changed across the networks and conditions helped to deduce a more logical value (Fig 6.7). From this analysis, it could be seen that there were several predicted interactions that were temperature sensitive. For example, the regulation of CCA1/LHY by GI had a high strength at 22˚C in BL, however had a weak strength in every other condition. Similarly, TOC1's interaction onto PRR7/9 appeared to reduce in strength as the temperature increased. This information was then also used to see how well the inferred network matched the Pokhilko et al. 2010 model at varying threshold values. This was done by comparing connections in the model to those in the inferred network. Any connection that appeared in both was scored as having distance 0,

A)



B)

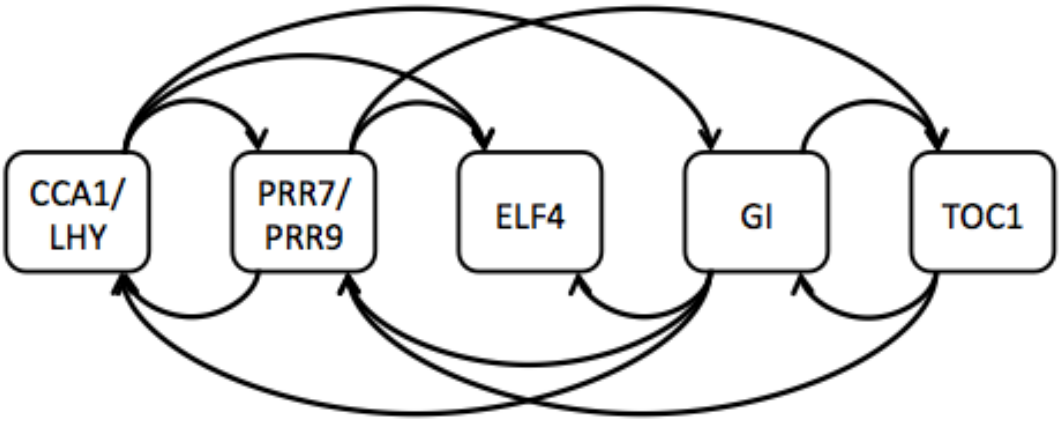


Figure 6.6 Visualisation of the inferred network from 22˚C BL compared to the Pokhilko et al. 2010 model. A) A modified image of the model created in Pokhilko et al. 2010. B) The top 10 connections of the network created by CSI, using data from the genes present in Pokhilko et al. 2010 plus ELF4.

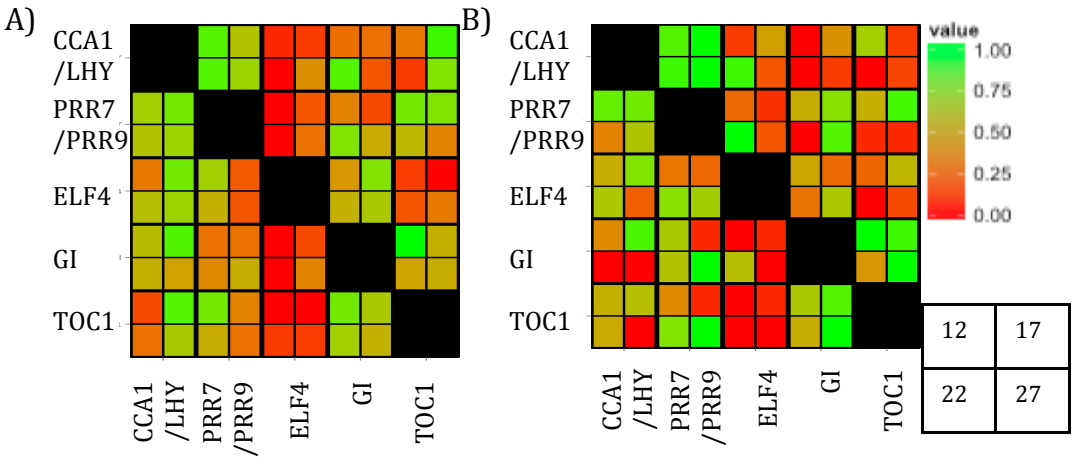


Figure 6.7 Heat maps of the strength (as shown in Supplemental Figure 6.1) of connections calculated by CSI for various conditions. CSI was run on the 6 genes named in the Pokhilko et al. 2010 model plus ELF4. Outputs for CCA1 and LHY as well as PRR7 and PRR9 were then combined. The resultant matrix was then colour coded where a value near 0 was coloured red and a value near 1 was coloured green. This was done for all four temperatures under A) red light and B) blue light. Each intersect between two genes is further divided into 4 boxes, depicting which temperature that strength relates to, as shown in the key.

and a connection that occurred in one but not the other was scored as having distance 1. By then summing up these scores (ignoring connections involving ELF4 which was not present in the model) a distance measure between the output of CSI and the model was calculated (Figure 6.8). In this figure, it was seen that under 17˚C, blue light conditions, CSI exactly matched the model when using a threshold of 0.31-0.51. Many of the other conditions also shared a minimum in their distance score around this point. Additionally, there was evidence that red light deviated more from the model than blue light did, as had previously been identified (Gould et al. 2013). These graphs also suggested that lower temperatures inferred luciferase networks were closer to the Pokhilko et al. (2010) model than the higher temperatures were. It was worth noting, however, that the Pokhilko et al. (2010) model was based on plants grown in white light conditions on media containing sucrose, this may impact the distance scores as the luciferase plants were grown under a single light colour with no sucrose in the media.

Using a threshold of 0.45 (chosen because several models had a reduced distance score at that threshold), the networks for the different temperature and light conditions were produced (Fig 6.9). From these networks, it could be identified that the morning (LHY/CCA1 and PRR7/9) and evening (GI and TOC1) loops were well conserved across the conditions. The way that these loops were connected, however, was greatly dependent on the conditions that generated the network. There is little evidence for this change in regulation within the literature. However, the missing evening complex genes may be a causative factor. Adding in these genes may produce a more consistent network across the temperature range.

6.5.2 – 14 Component (expanded) 2012 Model

Following the success of recreating the Pokhilko et al. 2010 model in the different condition sets, the next step was to see how well CSI coped with an extended 14 gene model (Pokhilko et al. 2012). These genes include all 11 genes

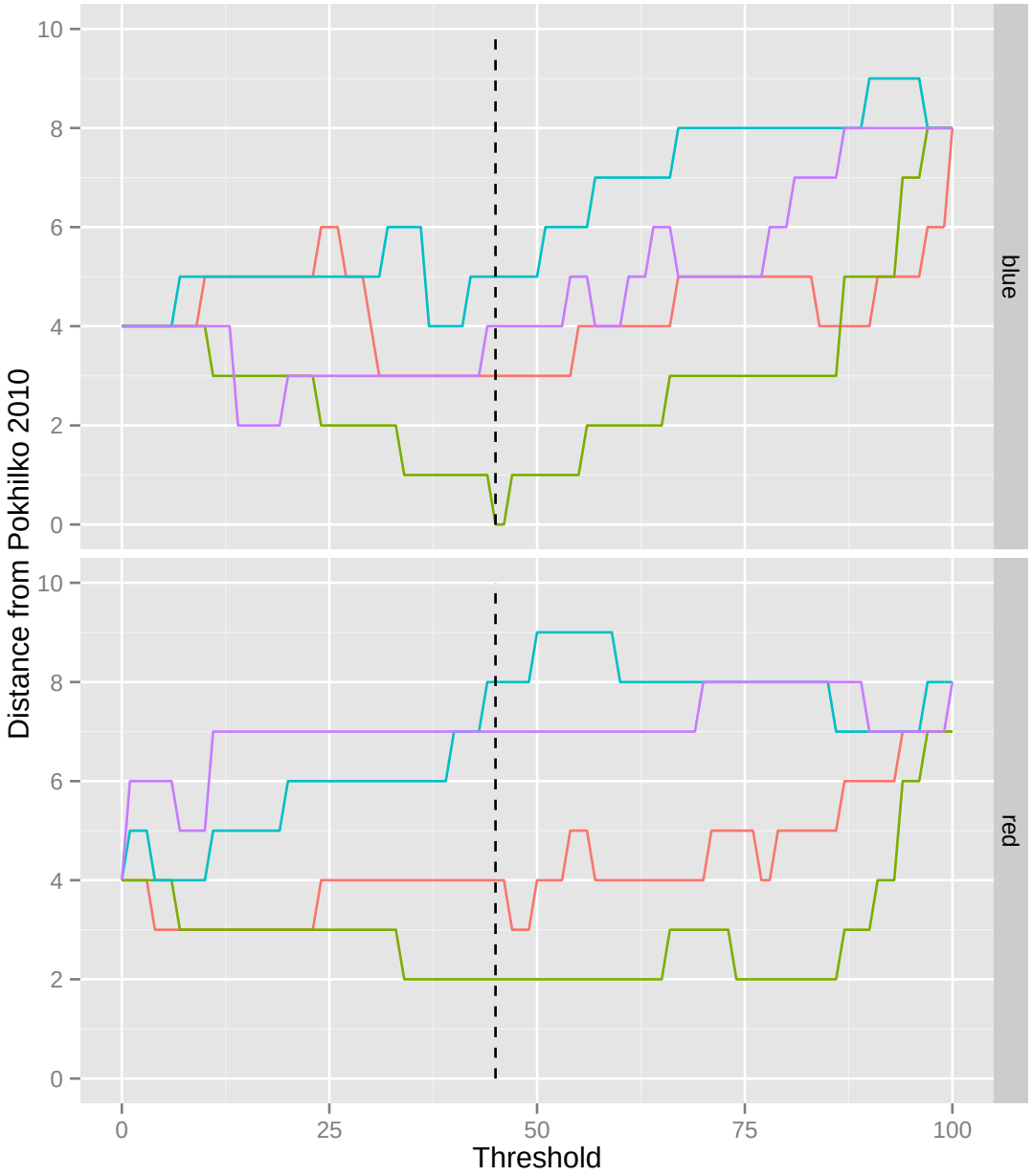


Figure 6.8 Distance between inferred luciferase network and the Pokhilko et al. 2010 model as the threshold was changed. Connections that were only present in one of the networks cause an increase to the distance score. This was done independently for each temperature and light condition calculating the score at various threshold values. Different temperatures are colour coded: 12˚C Orange, 17˚C green, 22˚C blue, 27˚C purple. Dashed line depicts the Threshold where CSI completely matches the Pokhilko et al. 2010 model for blue light at 17˚C.

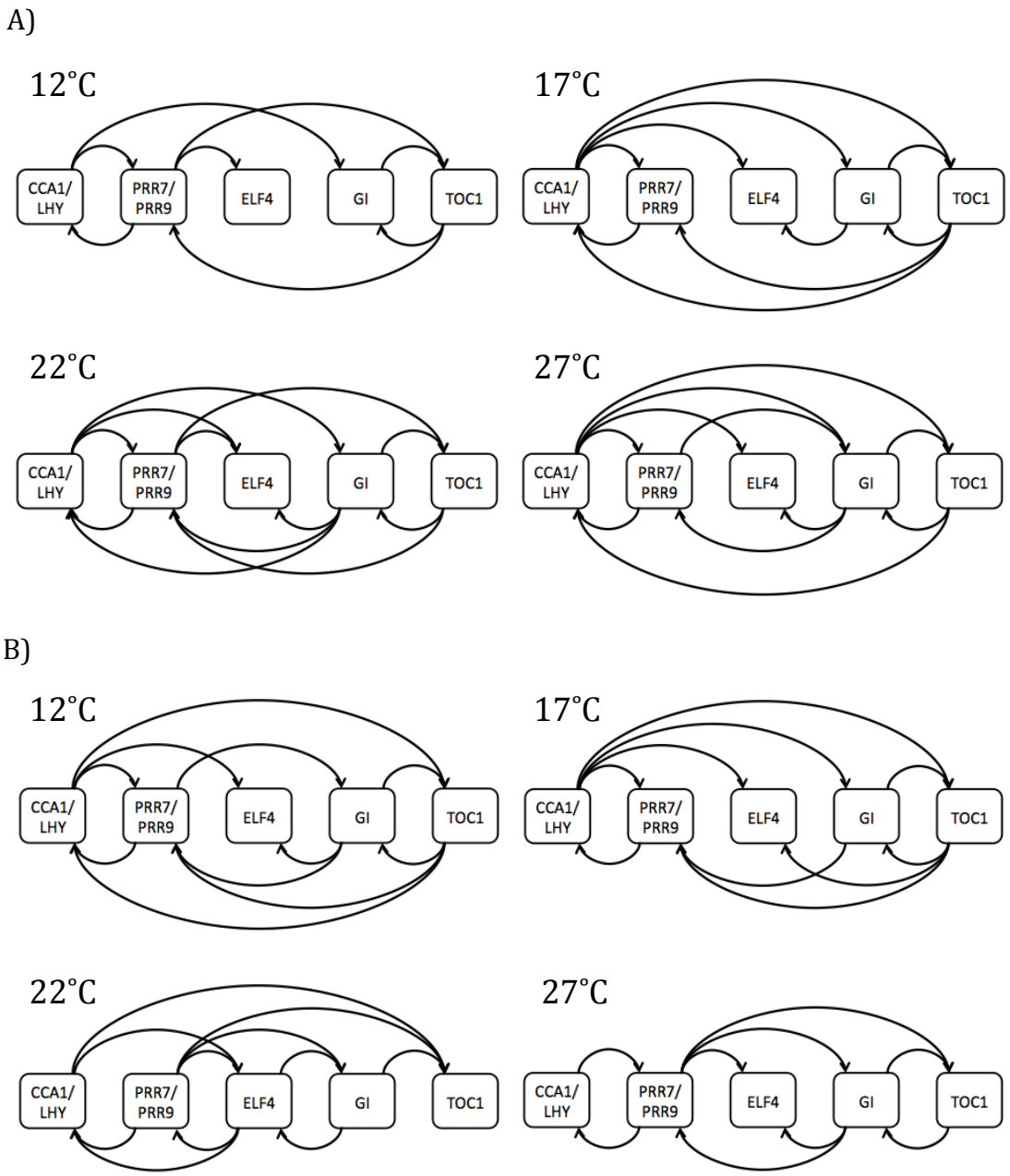


Figure 6.9 Inferred Networks from CSI using a threshold of 0.45. Data was produced using plants grown in A) blue light and B) red light in different ambient temperatures

named in the Pokhilko et al. 2012 model plus an additional clock gene TIC and two common reporter genes, CAB2 and CCR2. Even with this increased number of genes, the method of creating the pSet stayed the same (maximum of two simultaneous connections plus forced self interaction), as did the rest of the methodology used previously in the 7-gene network.

This analysis resulted in matrices similar to those used to inform tables 6.1 and 6.2 (Supplemental Table 6.1). Instead of combining genes to form the elements within this model, each gene was considered independently. These networks were then viewed in a similar way to the 7 gene networks, using a colour scale to show the relative strength of the different interactions (Fig 6.10). The most obvious result in this figure was that there was less variation in the colours, with most cells being red (i.e. very low connection strength scores). Since cMax was set at 2, assuming 1 pSet member was significantly better at explaining the data consistently across the repeats, the 2 connections within it would get a score of 1 and the other 11 would get a score of near 0. As more connections are likely to be real, the average score of those that are predicted decreases. The fact that many cells have a medium colour and very few are solid green suggested that there were more than two connections per gene that were needed to fully explain this increased network.

Due to this large number of predicted connections, analysing the output using a different method was performed. In Fig 6.5 and 6.6, networks were drawn using the strongest connections predicted by CSI. This method was tested using the network predicted under 17˚C blue light conditions (due to it having had the optimum result with the Pokhilko et al. 2010 model). Because of the larger number of genes, the top 28 connections were chosen. This number was chosen since during the runs of CSI, a gene's pSet was limited to two simultaneous regulators and as such 2 x 14 (number of genes) connections were expected (Fig 6.11). Looking at the inferred network connections, several of the connections within the Pokhilko et al. 2012 model could be identified, such as LHY feeding into the PRRs and evening genes or TOC1 feeding back onto the morning genes.

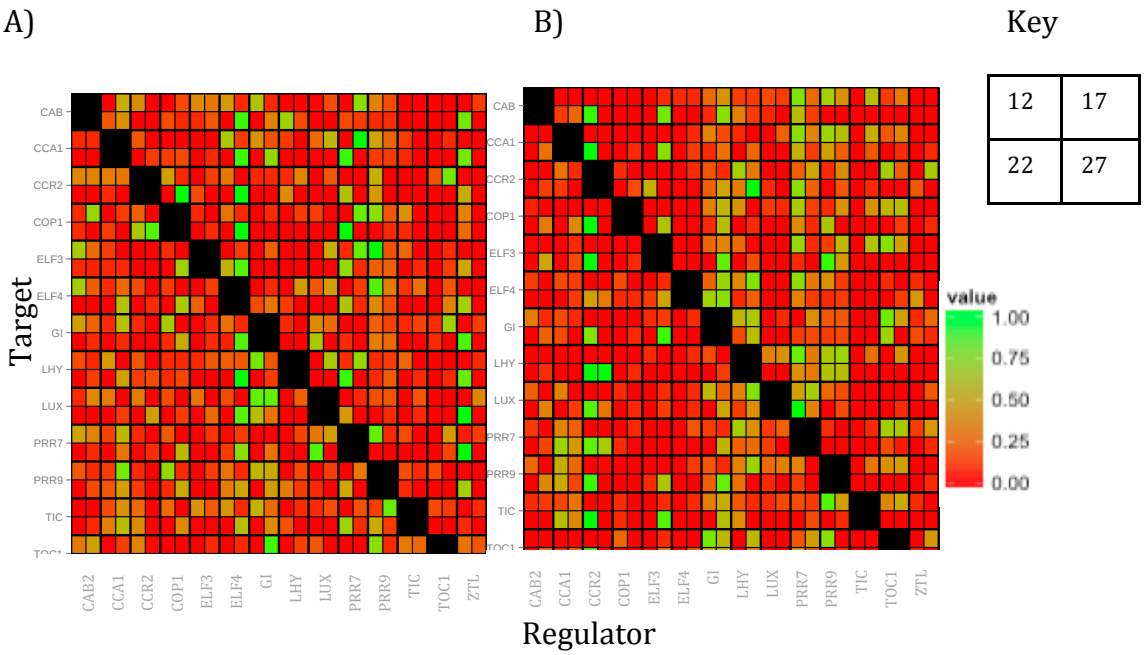
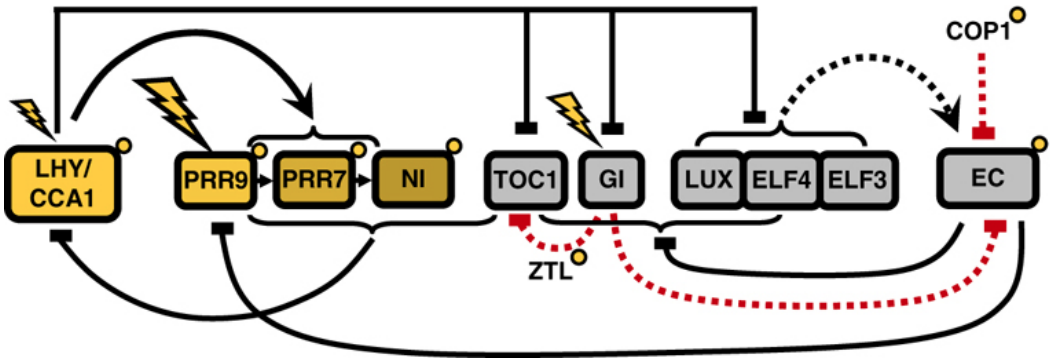


Figure 6.10 Heat maps of the strength of connections calculated by CSI for various conditions. CSI was run on the 11 genes named in the Pokhilko et al. 2012 model plus TIC, the outputs CAB2 and CCR2 were also included. The resultant matrix was then colour coded where a value near 0 was coloured red and a value near 1 was coloured green. This was done for all four temperatures under A) red light and B) blue light. Each intersect between two genes is further divided into 4 boxes, depicting which temperature that strength relates to, as shown in the key.

A)



B)

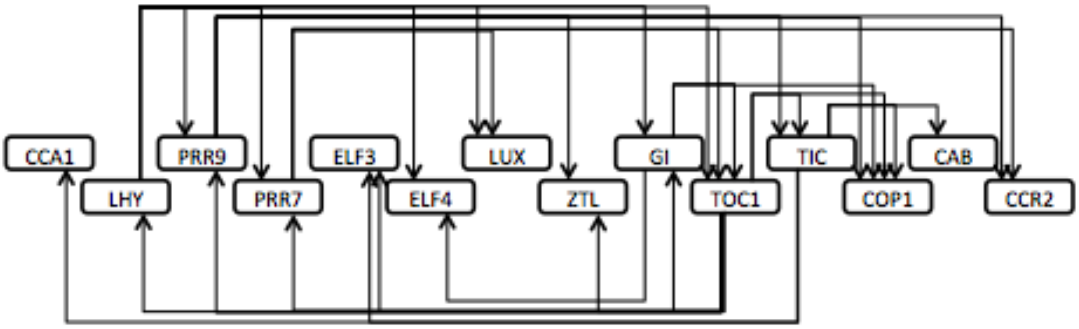


Figure 6.11 Visualisation of the inferred network from 17˚C BL compared to the Pokhilko et al. 2012 model. A) A modified image of the model created in Pokhilko et al. 2012. B) The top 28 connections of the network created by CSI, using data from the genes present in Pokhilko et al. 2012 plus TIC, CAB2 and CCR2.

However, within the inferred network there were no connections from members of the evening complex onto other genes. Similarly, although the PRR's did have an inferred role in regulating the evening genes (not found in the Pokhilko et al. 2012 model) they were not inferred to regulate LHY or CCA1 (which is in the Pokhilko et al. 2012 model). Running CSI on the other conditions produced similar results (Supplemental Figure 6.2). The low number of simultaneous connections that were allowed in the pSet could have caused this reduced efficiency in reconstructing the network. However, with the increased size of the network, increasing this value required more computational time than was permitted within the project. However, although the thresholded network retained very little of the Pokhilko et al. (2012) model, the entire matrix might have contained enough information to create a Boolean model as previously explored in Chaper 5.

6.6 – Modeling CSI Networks

To model the networks produced by CSI, the connection strengths being outputted needed to be identified as either an activator or an inhibitor. Once this feature had been added to CSI, the new connection matrix could be fed into the same simulation software previously made for VBSSM (Chapter 5.3). After simulations were optimised to produce stable oscillations, these models would be used to investigate how the circadian clock would change when specific genes were removed (knocked out) under the different conditions.

6.6.1 – Adding a Sign to the Connections

CSI investigated the ability of members in the pSet to accurately predict the expression of a gene by fitting an n-dimension plane to the data. This plane did not need to be flat, thus the interaction between two genes need not be linear. Because of this, choosing how to determine the sign of the interaction was crucial. The point where the fitted plane was most likely to describe the direct interaction of the genes was when expression of the gene causing the interaction was high. Additionally, there needed to be sufficient data at that point to be confident of the slope of the plane when it was measured. Due to dampening rhythms, this made the peak of expression a poor choice as it excluded many of the oscillations (Fig 6.12). Averaging the last peak did not always work either, especially when there was little dampening or when the expression became arrhythmic. As such averaging all the expression values for a gene produced the most reliable point for determining the sign of the connection being inferred. The gradient at this midpoint could then be calculated by measuring the instantaneous change in levels at that point, the sign of this gradient was exported and used to determine activation or repression. Sometimes the direction of the gradient of an interaction varied between the pSet members. In those cases, the sign of the interaction in the strongest pSet member was used to determine the sign of a genes interaction. After adding this into the code, CSI was rerun for the 14 gene networks.

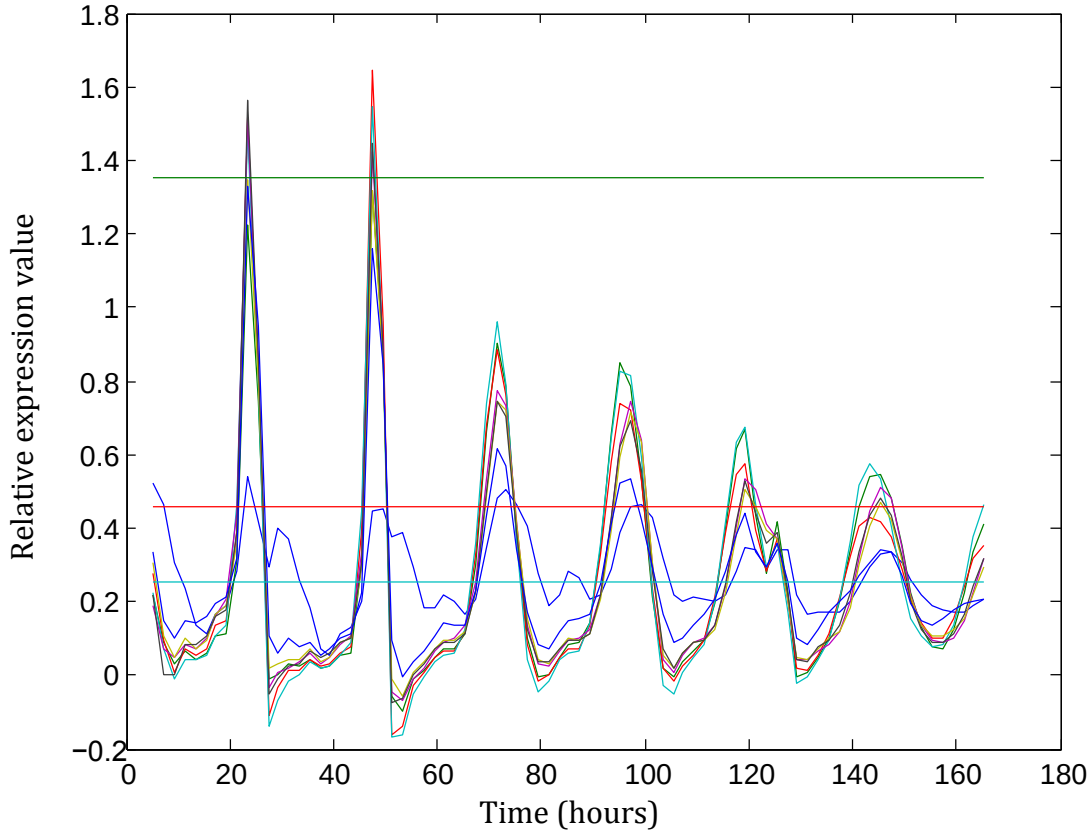


Figure 6.12 Plot of expression data for CAB2 and intersect of possible sign inference points. The top (green) horizontal line represents the average of the first peak. The middle (red) line represents the average of the last peak. The bottom (blue) line represents the average of the entire dataset.

6.6.2 – Boolean Modeling

Using the new output of CSI, probabilistic Boolean modeling was run to see if the matrix outputted by CSI was capable of producing oscillating graphs. Applying the same analysis to optimise conditions as previously explored, (Chapter 5.3) simulations with the BL 22˚C data failed to produce stable oscillations using any of the start states or incMax values. Applying a threshold to connections (similar to that chosen for VBSSM) also failed to produce stable oscillations in the simulations. Additional attempts to scale connections, such as log scaling to reverse the exponential scaling that occurred in CSI before data was outputted, also failed to create the desired simulation results.

This failure may be an artifact of the inner workings of CSI. Because each gene was evaluated for regulators independently, they were also scaled independently. This meant that each gene in the model was scaled to the same total regulation, even if that was not the case in the biological system. Additionally, the value being outputted is a statistic of how tight the data was to the fitted plane, corrected based on how complex that plane was. As such it was not a measure of how much the connection affected the gene, but rather how well the genes expression could be predicted using that connection. Thus, modeling the output was potentially not possible without large changes to fundamental aspects of how CSI analysed the data.

6.6.3 – Analysing Effect of Sign

As seen above, the addition of the sign was not sufficient to allow the output to be modeled using the framework created in Chapter 5. However this could have been because the method of determining the sign was inefficient. To test this, a similar distance analysis to that which generated Figure 6.8 was applied. To expand the distance score, a match in connections between the inferred network and the Pokhilko et al. 2010 model was given a distance 0, a connection in one network but not the other was given a distance 1, and a mismatch in the

signs was given a distance 2. To test this in a controlled manner, the 7-gene model previously used in Chapter 6.5.1 was rerun and the interaction signs were obtained. Previously, when calculating the union of two genes to form an element in the model, the highest strength connection was retained. However, now that each connection had a sign as well as a strength score, it became more complex when the two strengths being combined were of opposite signs. When there was a large difference between the scores, it was not that important and the largest value could still be used as previously done. However, when the difference in the absolute values was relatively small, it became more difficult to justify which one to retain. Additionally, there were some cases where the absolute values were identical but had opposite signs (PRR7/9 interaction onto ELF4). Fortunately, this did not occur in connections that were within the Pokhilko et al. 2010 model. As such, these connections that had opposite signs of equal strength (e.g. PRR7/9 interacting onto ELF4, Supplemental Table 6.2) could be combined by using the sign recovered in the networks inferred at other temperatures (Fig 6.13). In this figure, several differences between the difference distances could be seen. The previous graph with a distance of 0, seen under blue light 17˚C conditions, no longer reached 0. Instead it reached a minimum at a distance of 4, caused by two connections that were predicted to have the opposite sign than was seen in the model. This was even more apparent at 27˚C blue light, where virtually every connection was inferred to have the opposite sign, with a maxing distance of 12. There were also large differences between red and blue light data and how they changed when the sign of the interaction was calculated. This suggested that not only did the importance of different genes change at various temperature and light conditions, but also the type of effect they have within the network. Within this investigation, there were no instances where the sign of every connection matched that in the Pokhilko et al. 2010 model. However, this did not necessarily mean that the method was flawed. The number of connections where the sign predicted matched the sign described in the Pokhilko 2010 model (43/64) was significantly higher than would be expected if the sign recovered was random (32/64). Additionally, some of the connections within

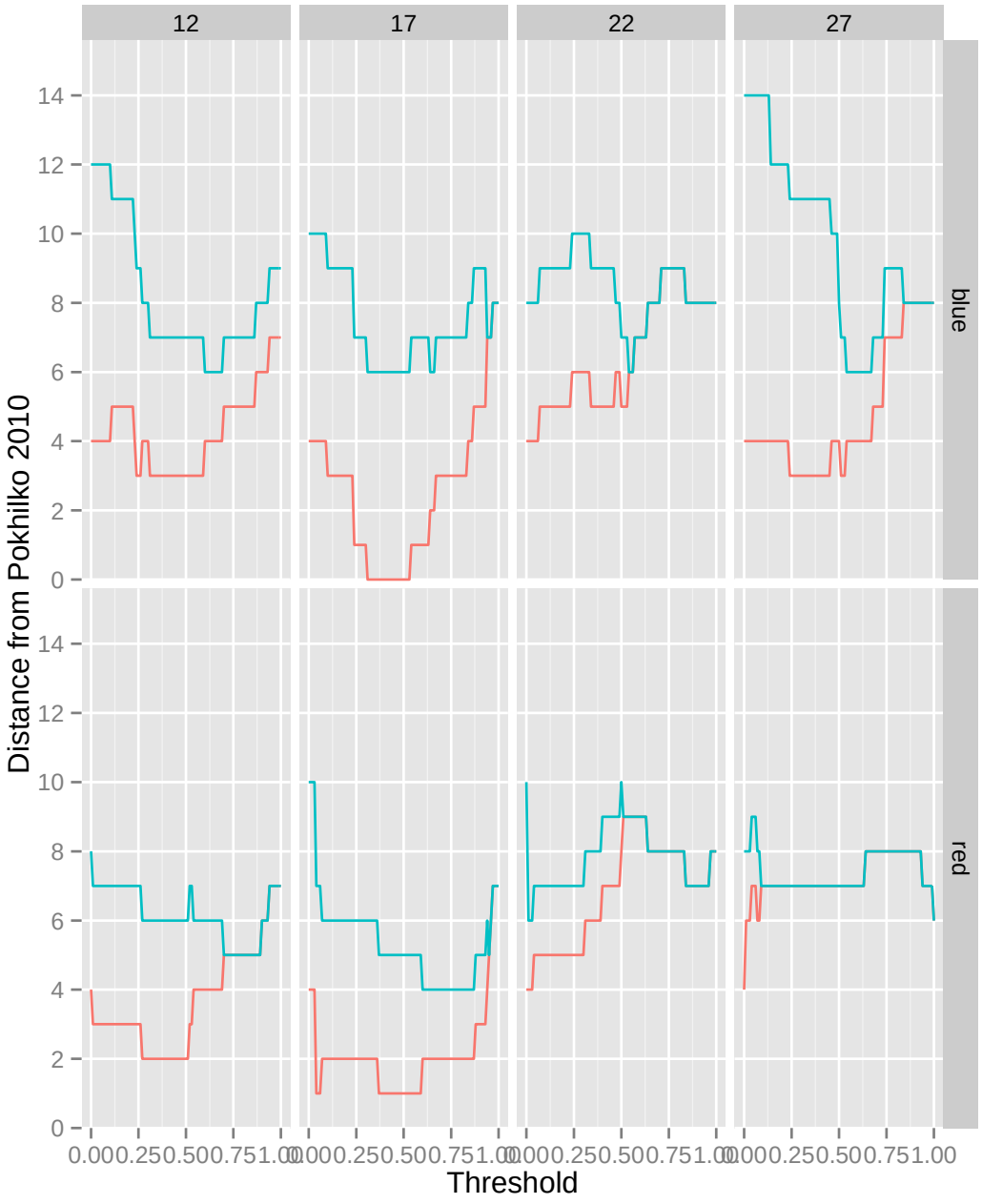


Figure 6.13 Distance between inferred network and the Pokhilko et al. 2010 model as the threshold was changed. Connections that were only present in one of the networks cause an increase to the distance score by 1. Connections where the sign is different get a score of 2. This was done independently for each temperature (column) and light condition (row). Distance score using the adapted CSI (blue) is plotted alongside the distance score without using the sign (orange).

the Pokhilko et al. 2010 model may be wrong. Also of note was the fact that the signs that were predicted as opposite to the model were not consistent across all of the networks tested, meaning that these differences are not an inherent disagreement with the model, but rather a potential method of temperature compensation within the clock.

6.7 – Discussion

CSI was a network inference software package similar to VBSSM. However, the way CSI calculated the strength of interactions meant that it was better able to cope with the cyclic nature of circadian genes: when the code was run multiple times with different seeds, it still returned the same network. That said, it was not designed for luciferase data, and as such it was sensitive to the order the data was inputted in. Running the code multiple times with randomly mixed luciferase datasets converged to a single network. This need to run CSI multiple times was also likely unneeded if enough random pairing of genes were submitted within a single simulation.

When simulated data was processed by CSI, it produced a network that roughly matched the model that created the data. However the lack of protein data hampered the inference of specific connections. Additionally, the model did not include mRNA data for all the components modeled in the simulation, limiting the ability to capture all the connections. When the equivalent luciferase data was passed into the software, there were some similarities between the luciferase inferred network and the network produced using simulated data. However there were also considerable differences, including more connections in the luciferase network. This suggested that the model, although capable of matching a lot of the data, still did not include all of the interactions that were within the plant. Additionally, the connections that were recovered in both simulated and luciferase data sets tended to be of less importance within the luciferase networks. Thus, CSI should be used to create novel networks rather than building on top of an existing one. This was predicted to provide more interesting and realistic results.

When CSI was used to infer the network behind the Pokhilko et al. 2010 model components, it produced an identical network at 17˚C in blue light. However the Pokhilko et al. 2010 model was based at 22˚C in white light. Other temperature conditions and red light conditions produce different networks. This showed the network needed additional components to explain why the topology

changed so much. The network was expanded to include the 11 genes modeled by Pokhilko et al. 2012, as well as a few additional genes. This process inferred another set of networks under the different conditions. These networks loosely fit what was predicted in the model, with elements most similar to those in the reduced Pokhilko et al. 2010 model being recovered more frequently. The resulting networks, however, lacked information about connections currently modeled as having a strong protein modification step.

Despite this lack of protein data, CSI still returned connections involving the individual genes of the protein complex and it's eventual downstream target. This showed that whilst CSI may not recover all the connections predicted by the model, if the causality between them is high enough it will be identified despite the missing intermediary connections. This provided a potentially powerful tool for analysing data to identify possible paths of regulation. As the networks contained many interconnected loops, an informative method to examine the output of CSI without reducing the information produced by CSI would be to model the network. This would provide a more meaningful interface to test how sensitive each theoretical connection was to changes. To achieve this, however, additional information about the inferred network needed to be generated by CSI. CSI lacked the capability to describe the connection it predicted; it just said how well that connection explained the expression of a gene. Adding in the ability to detect the type of interaction being predicted (promotion or repression) was also insufficient to provide enough information to generate a model that generated oscillating simulations. Additional adaptions to CSI may make it possible to model the network. Alternatively rerunning CSI, but increasing the number of simultaneous regulations permitted would change the heat maps produced above so that the range of values are better distributed.

Comparing the results using the adapted version of CSI to the Pokhilko et al. 2010 model showed a large number of connections had the same sign as was modeled for the different temperature and light conditions, although individual connections often changed across this condition range. There are several

deviations, however, the number of which depends what conditions the network was being studied in. The connections that have a sign opposite to the model vary across the conditions. This suggests that not only do the different connections have a variable importance under the conditions within the condition set; the connections can have different interactions.

CSI has been shown to reliable produce a network with large similarities to existing models. Using simulated data from the Pokhilko et al. (2012), CSI recovered several of the connections that describe this model. However, it does not fully recreate the network. Much of this can be explained through missing protein information, and the limit to the number of simultaneous interactions. CSI's ability to recover this network is therefore severely reduced. The Pokhilko et al (2010) model has far fewer protein modification steps, and any one component has more than 2 regulators. Using simulated data from this model may better determine CSI's usefulness to reproduce circadian networks before time is spent generating protein information or running CSI with a greater number of members in the pSet.

Chapter 7 – Final Discussion

7.1 – Introduction

With the increase in high throughput experiments, there is now a major issue in dealing with large datasets. As experimental techniques develop, a single experiment produces information on more aspects of the system being analysed, and more detail on each aspect. This becomes especially problematic when trying to understand and identify important components in the system being investigated. Larger datasets may allow better constraints to the model, but the additional components exponentially increase the time needed to optimise models to test whether connection x or y is key to understanding the system.

An example of this is the circadian clock within Arabidopsis. Over the last decade it has expanded from a simple 3-gene feedback loop (Alabadí et al. 2001) to a more complex 11-gene repressilator (Pokhilko et al. 2012). Despite this, many experimentally proven mechanisms are not included in modeling (such as LHY and CCA1 acting independently of each other) and many genes with severe circadian phenotypes are not within the model (e.g. TIC). Additionally, current simulations are based on a subset of data taken from specific conditions (22˚C white light data from plants grown on 3% sucrose). Expanding these models requires the addition of connections and components, as well as the optimisation of an increasing number of variables. Many of these models will not produce a better simulation, however, so alternative models then need to be developed, optimised, and tested. This requires considerable time and computer resources, finding a way to direct models without needing to perform this optimisation of each possible model will greatly improve efficiency.

7.2 – Microarray Analysis

Initially, existing bioinformatics tools were used to examine microarray data performed at several temperatures in wild type and a temperature compensation mutant. These results showed that temperature had an effect on a large scale, with transcription, translation, and protein degradation having significant differential regulation within the temperature range. The results also showed that the *gi-11* mutant had similar temperature dependent phenotypes. However, the scale of the differential expression was significantly increased in the mutant background. This effect of the *gi-11* null mutant suggested that although the circadian clock has evolved a strong temperature compensation mechanism, the plant as a whole is buffered to temperature through a mechanism separate from the circadian clock. Thus, even when temperature compensation within the clock is disrupted, general reactions to a change in temperature persist. This analysis further supported the theory that transcription and translation within the circadian system is a strong candidate for temperature compensation (Sideaway-Lee et al. 2014).

In addition, delayed fluorescence screening identified 13 genes that had a temperature specific period effect. Within this list were two known core circadian clock genes, PRR5 and PRR3. These were the only circadian clock genes with a two-fold expression change in response to temperature and a differential expression caused by the *gi-11* mutation. They have previously been shown to have a minor role in sensing themocycles (Eckardt 2005) but are not part of current circadian models (Pokhilko et al. 2012). Understanding how these genes fit into the circadian network, and how their expression is linked to temperature, could provide the next development to circadian models.

In addition to these known clock genes, this screen identified multiple genes which have been recorded as having roles in salt stress response (Li et al. 2013; He et al. 2005). This system has been published as having links with both temperature sensing and photosynthesis, potentially linking the reporter system used in delayed fluorescence and the temperature compensation that it

was used to investigate. This DF screen also identified several genes linked to the chloroplast, potentially targets for understanding how the circadian clock is coupled to Photo System II. This linkage is crucial to understanding how this inbuilt circadian reporter can be used to truly monitor what is happening to the circadian clock.

7.3 – Visualising Multiple Cluster Sets

Using luciferase data, the circadian clock was then examined for genes that were similarly expressed across the temperature range. Whilst there were several cluster methods that could cluster the data from a single condition, due to the richness of the data sets, existing software failed to cluster genes when multiple conditions were used to simultaneously inform the clusters. As such, new software was designed to display how the genes clustered in each of the conditions separately. Using this new tool, several genes were identified as having a similar expression pattern in all of the conditions. Additionally, some genes were identified as being similarly expressed in some conditions, but expressed differently at others.

This included the co-clustering of CCA1 and LHY, two genes in the morning loop that are frequently modeled as a single element (Locke et al. 2006, Pokhilko et al. 2010, Pokhilko et al. 2012). This co-cluster occurred at every temperature and light condition. This suggested that this method was capable of recovering meaningful data. In addition to this and other known clusters, this method identified several genes that were frequently clustered together. Most significant of these was ELF3 and SPA3. These genes are involved in the evening complex and light input respectively. Thus, this suggests a link between light sensing and a major component of the evening loop. Other elements of the evening loop are modelled as having a light dependency, so this hypothesises an increased control of the evening loop by light.

Higher throughput experimental designs have become more common, with a large number of elements (e.g. genes) being measured for a range of different phenotypes. However, many of those different phenotypes were not readily comparable to each other. The use of n-dimensional clustering could also be difficult when dealing with long, high resolution time series data or equivalent. Existing software that tried to cluster the elements on multiple of these distinct

experiments were often limited by the type of data they are able to accept. Also, there was seldom a way to select specific cluster methods for each data set.

With the newly developed software, elements can be clustered or categorised for different phenotypes independently. The groups created are then plotted on top of each other. This enables a visual comparison between which elements were grouped together under the different conditions. Within this thesis, this software was created and improved using the cluster results of luciferase expression of different genes. However it was designed in such a way that other datasets can be handled equally as well. Thus elements could be visualised not only on how cluster software splits them, but on other quantitative and qualitative traits as well. As such this software could be useful in a wide range of different disciplines.

One example is competition studies of light and temperature cycles on the circadian clock. Both temperature and light changes can reset the circadian clock (Harmer 2000; Salome et al. 2005). However little is known about how this signal would combine if they were set antiphase to each other. Experiments have attempted to identify whether genes are more sensitive to one than the other, however a consensus clustering method would provide a quick overview of this result. Generating time series for genes under all three conditions (light cycles only, temperature cycles only, both light and temperature cycles but antiphase to each other), each set of results can then be clustered independently. By then plotting these clusters on top of each other, genes can be investigated for how there cluster membership is affected. This will then generate lists of genes that are only sensitive to one or other signal, as well as a list of genes sensitive to both.

7.4 – Use of Network Inference to Build the Circadian Network

Whilst both microarray and cluster analyse provided an insight into how the circadian network changed, they did not suggest a direct way to restructure the clock to include temperature compensation. As such, a subset of the luciferase data was used to create networks using several network inference software packages. These packages were shown to be able to produce networks similar to those published when using the equivalent data set. Additionally, it was seen that these methods could be used to expand the network to include additional components. When the network inference was done on the separate temperatures, networks were produced which significantly differed from each other. Whilst general topology was conserved, the specifics of each network varied between the conditions. These variations in the inferred networks showed sections where the clock's architecture was most susceptible to temperature changes.

The creation of models has always centred on creating a network and then optimising the variables in order to match major dynamics of measured data. However choosing which components to include in the network and how they interact requires a lot of experimental work. Even then, it often requires trial and error to determine which connections are needed, with each possible network being optimised and simulated. Using time course data to calculate the interaction between different components has been performed for several years, although it has been mostly confined to smaller, linear signalling pathways.

By using existing software, luciferase data was shown to be capable of recovering a network very similar to an existing model of the circadian clock. When this method was extended to consider the components of a more complex model, many of the connections were still recovered despite the lack of protein data. These networks could therefore have been used to develop the model in a

similar manner to the original development. By then using more genes from the original dataset, the network can be further expanded.

The raw outputs of these methods, however, could not be used to create a working probabilistic Boolean model. This was caused by a limitation of the code to produce consistent measures of interaction strength in the case of VBSSM. When CSI was used, it consistently predicted a set of connections with similar connection strengths between repeats. However the output values had no sign of interaction. Indeed even when the sign of interaction was added, CSI was still unable to create oscillating simulations. This was caused by the outputted statistic being a measure of how well the fitted plane captured the data, but gave no information about the plane. Adapting the code to provide information about the gradient of the plane may succeed in producing enough information to inform a simple simulation. Alternatively, if the code cannot be adapted to produce oscillating simulations in it's own right; the networks inferred can be used to guide the construction of other models by providing a list of gene interactions predicted to occur.

CSI did highlight several interesting features, however. Firstly, it predicted that LHY/CCA1 promoted TOC1. This is in direct conflict with existing models, however the evidence for this connection is based on constitutive overexpression or knock out of the LHY or CCA1 gene and a TOC1 luciferase reporter. However, the same style of information was used to predict TOC1 promoted LHY/CCA1, which has since been found to be a repressive connection. Using inducible over expressers, this connection can be explored in more detail to discover whether CSI has uncovered new information, or whether it has incorrectly predicted the sign. Should CSI be proved right, there are addition connections that could be analysed in this way, some of which are predicted to be temperature specific.

7.5 – Future Perspective

Many aspects of this work can be built on further. The initial finding of GO terms differentially expressed across the temperature range suggested a role of the cell wall within temperature compensation. This modification of growth within plants could be important, especially in light of transcriptionless oscillators. These oscillators alone can function as a circadian clock, but only as long as cell division is longer than a 24 hour period (Zwicker et al. 2010). Additionally, the antagonistic responses of transcription, translation and protein degradation as ambient temperature changes could be explored in greater detail; does the inhibition of one of these processes have a greater effect at different temperatures. Use of chemicals to inhibit either the transcription or translation process could prove this hypothesis.

Additionally, delayed fluorescence screens identified several ways the circadian network should be expanded to model temperature changes and incorporate the DF reporter system. Verifying these findings with QPCR and further investigation of the individual genes may result in a model of the circadian network, buffered against temperature, with connections to an inbuilt reporter system. This expanded model would act as a powerful tool in predicting, and testing, how the plant would respond to different stimuli or gene mutations.

Through the consensus clustering investigation, several sets of genes were identified as being co-expressed in a range of conditions. Experiments into how these genes are regulated may reveal why they were being co-expressed across the different conditions. Observing downstream factors and the effect of single and double knockouts may also identify why they are being co-expressed. This clustering method is also designed to accept additional ways genes can be grouped. This includes things like whether they have specific binding sites on their promoter, or how much a specific gene mutation effects expression. By using different data sets, genes can be explored for how they compare in a range of conditions.

Use of network inference software can be used to rapidly expand the components used to model the circadian clock. Within this project only a small subset of the luciferase data available was used to infer networks. This subset did include an additional core oscillator as well as a couple of clock output genes. It did not, however, include light input components (such as CRY or PHY) or a range of hypothesised circadian components (like PRR3). Also, the differences between the networks produced at the various temperatures can be further investigated to understand why the topology has changed so much. Much of this is likely due to the limited number of simultaneous connections, which artificially reduces the degree of interaction in the network. Rerunning with a higher cMax value would fix this; however there would still be a difference in gene importance in explaining a genes expression. The most efficient way to further investigate this work would be to simulate the different networks. This would provide a framework to form testable hypotheses concerning knock out mutants. With this validation of the method, increasing cMax and the number of genes being used to form networks can be undertaken to create a more exhaustive circadian network, which contains the ability to adapt to environmental changes.

Adapting CSI to also produce information on the predicted type of connection being predicted produced hypotheses that some connections changed from promoters to inhibiters with temperature. It also predicted that the LHY/CCA1 repression of TOC1 present in all existing models should actually be a promotion. Using a mutant with an inducible gene could test these predictions. For example, by using a wt plant with an inducible LHY construct, TOC1 could be screened for immediate (or slightly delayed) responses to increased LHY levels. This would determine exactly how TOC1 behaves to the presence of LHY.

Chapter 8 – Bibliography

Akman, O.E. et al., 2012. Digital clocks: simple Boolean models can quantitatively describe circadian systems. *Journal of The Royal Society Interface*, 9(74), pp.2365–2382.

Alabadí, D. et al., 2001. Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science*, 293(5531), pp.880–883.

Alabadí, D. et al., 2002. Critical Role for CCA1 and LHY in Maintaining Circadian Rhythmicity in Arabidopsis. *Cell*, 12(9), pp.757–761.

Alonso, J.M., 2003. Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana. *Science*, 301(5633), pp.653–657.

Aschoff, J., 1963. Comparative physiology: diurnal rhythms. *Annual review of physiology*, 25, pp.581–600.

Bayer, R.G. et al., 2011. Mining the soluble chloroplast proteome by affinity chromatography. *Statistics in Medicine*, 11(7), pp.1287–1299.

Beal, M.J. et al., 2005. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3), pp.349–356.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.289–300.

Botstein, D. et al., 2000. Gene Ontology: tool for the unification of biology. *Nature*, 25(1), pp.25–29.

Breeze, E. et al., 2011. High-Resolution Temporal Profiling of Transcripts during Arabidopsis Leaf Senescence Reveals a Distinct Chronology of Processes and Regulation. *The Plant cell*, 23(3), pp.873–894.

Camon, E., 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(90001), pp.262D–266.

Carpenter, C.D., Kreps, J.A. & Simon, A.E., 1994. Genes encoding glycine-rich Arabidopsis thaliana proteins with RNA-binding motifs are influenced by cold treatment and an endogenous circadian rhythm. *PLANT PHYSIOLOGY*, 104(3), pp.1015–1025.

Chinnusamy, V., Zhu, J. & Zhu, J.-K., 2007. Cold stress regulation of gene expression in plants. *Trends in Plant Science*, 12(10), pp.444–451.

Cosgrove, D.J., 2005. Growth of the plant cell wall. *Nature*, 6(11), pp.850–861.

Costa, M.J. et al., 2013. Inference on periodicity of circadian time series. *Biostatistics (Oxford, England)*.

Costa, M.J. et al., 2014 ReTrOS: a tool for reconstructing the temporal profile of transcription from imaging time-series. In press

Crosthwaite, S.K., Dunlap, J.C. & Loros, J.J., 1997. Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of circadian rhythmicity. *Science*, 276(5313), pp.763–769.

Dalchau, N. et al., 2011. The circadian oscillator gene GIGANTEA mediates a long-term response of the Arabidopsis thaliana circadian clock to sucrose. *Proceedings of the National Academy of Sciences*, 108(12), pp.5104–5109.

Desfeux, C., Clough, S.J. & Bent, A.F., 2000. Female reproductive tissues are the primary target of Agrobacterium-mediated transformation by the Arabidopsis floral-dip method. *PLANT PHYSIOLOGY*, 123(3), pp.895–904.

Diernfellner, A. et al., 2007. Long and short isoforms of Neurospora clock protein FRQ support temperature-compensated circadian rhythms. *FEBS Letters*, 581(30), pp.5759–5764.

Dodd, A.N., 2005. Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage. *Science*, 309(5734), pp.630–633.

Domijan et al., 2014. SASSy: Software for Analysis of Sensitivity of Systems. In press.

Downton, W.J.S., Grant, W.J.R. & Robinson, S.P., Photosynthetic and Stomatal Responses of Spinach Leaves to Salt Stress. *plantphysiol.org*.

Eckardt, N.A., 2005. Temperature entrainment of the Arabidopsis circadian clock. *The Plant cell*, 17(3), pp.645–647.

Edwards, K.D., 2006. FLOWERING LOCUS C Mediates Natural Variation in the High-Temperature Response of the Arabidopsis Circadian Clock. *THE PLANT CELL ONLINE*, 18(3), pp.639–650.

Edwards, K.D. & Millar, A.J., 2007. Analysis of circadian leaf movement rhythms in Arabidopsis thaliana. *Methods in molecular biology (Clifton, N.J.)*, 362, pp.103–113.

Eriksson, M.E. et al., 2003. Response regulator homologues have complementary, light-dependent functions in the Arabidopsis circadian clock. *Planta*, 218(1), pp.159–162.

Feldman, J.F. & Hoyle, M.N., 1973. Isolation of circadian clock mutants of Neurospora crassa. *Genetics*, 75(4), pp.605–613.

Fowler, S. et al., 1999. GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with several possible membrane-spanning domains. *The EMBO journal*, 18(17), pp.4679–4688.

Fry, F.E.J., 1958. Temperature Compensation. *Annual review of physiology*, 20(1), pp.207–224.

Garceau, N.Y. et al., 1997. Alternative Initiation of Translation and Time-Specific Phosphorylation Yield Multiple Forms of the Essential Clock Protein FREQUENCY. *Cell*, 89(3), pp.469–476.

Gautier, L. et al., 2004. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3), pp.307–315.

Gendron, J.M. et al., 2012. Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. *Proceedings of the National Academy of Sciences*, 109(8), pp.3167–3172.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Goodwin, B.C., 1966. An Entrainment Model for Timed Enzyme Syntheses in Bacteria. *Nature*, 209(5022), pp.479–481.

Gould, P.D. et al., 2006. The Molecular Basis of Temperature Compensation in the Arabidopsis Circadian Clock. *THE PLANT CELL ONLINE*, 18(5), pp.1177–1187.

Gould, P.D. et al., 2009. Delayed fluorescence as a universal tool for the measurement of circadian rhythms in higher plants. *The Plant journal : for cell and molecular biology*, 58(5), pp.893–901.

Gould, P.D. et al., 2013. Network balance via CRY signalling controls the Arabidopsis circadian clock over ambient temperatures. *Molecular systems biology*, 9, pp.1–13.

Hall, A., 2003. The TIME FOR COFFEE Gene Maintains the Amplitude and Timing of Arabidopsis Circadian Clocks. *THE PLANT CELL ONLINE*, 15(11), pp.2719–2729.

Harmer, S.L., 2000. Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. *Science*, 290(5499), pp.2110–2113.

Hartigan, J.A. & Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pp.100–108.

He, X.-J. et al., 2005. AtNAC2, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development. *The Plant Journal*, 44(6), pp.903–916.

Heard, N.A., Holmes, C.C. & Stephens, D.A., 2006. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes. *Journal of the American Statistical Association*, 101(473), pp.18–29.

Highkin, H.R. & Hanson, J.B., 1954. Possible Interaction between Light-dark Cycles and Endogenous Daily Rhythms on the Growth of Tomato Plants. *PLANT PHYSIOLOGY*, 29(3), pp.301–302.

Hill, A.V., 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J physiol*, 40(4), pp.iv–vii.

Hosokawa, N. et al., 2011. Circadian transcriptional regulation by the posttranslational oscillator without de novo clock gene expression in Synechococcus. *Proceedings of the National Academy of Sciences*, 108(37), pp.15396–15401.

Huang, Z.J., Curtin, K.D. & Rosbash, M., 1995. PER protein interactions and temperature compensation of a circadian clock in Drosophila. *Science*, 267(5201), pp.1169–1172.

Ishiura, M. et al., 1998. Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science*, 281(5382), pp.1519–1523.

Jain, A.K., Murty, M.N. & Flynn, P.J., 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3), pp.264–323.

James, A.B. et al., 2012. Alternative Splicing Mediates Responses of the Arabidopsis Circadian Clock to Temperature Changes. *THE PLANT CELL ONLINE*, 24(3), pp.961–981.

Kim, W.-Y. et al., 2007. ZEITLUPE is a circadian photoreceptor stabilized by GIGANTEA in blue light. *Nature*, 449(7160), pp.356–360.

Kitano, H., 2002. Systems biology: a brief overview. *Science*, 295(5560), pp.1662–1664.

Knight, M.R. & Knight, H., 2012. Low-temperature perception leading to gene expression and cold tolerance in higher plants. *Statistics in Medicine*, 195(4), pp.737–751.

Kumar, S.V. & Wigge, P.A., 2010. H2A.Z-Containing Nucleosomes Mediate the Thermosensory Response in Arabidopsis. *Cell*, 140(1), pp.136–147.

Kusakina, J., Gould, P.D. & Hall, A., 2013. A fast circadian clock at high temperatures is a conserved feature across Arabidopsis accessions and likely to be important for vegetative yield. *Plant, cell & environment*.

Li, W. et al., 2013. A Bi-Functional Xyloglucan Galactosyltransferase Is an Indispensable Salt Stress Tolerance Determinant in Arabidopsis. *Molecular Plant*, 6(4), pp.1344–1354.

Liu, D.W. & Thomas, J.H., 1994. Regulation of a periodic motor program in C. elegans. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 14(4), pp.1953–1962.

Liverani, S. et al., 2009. Efficient Utility-based Clustering over High Dimensional Partition Spaces. *Bayesian Analysis*, 4(3), pp.191–200.

Lobell, D.B., Schlenker, W. & Costa-Roberts, J., 2011. Climate Trends and Global Crop Production Since 1980. *Science*, 333(6042), pp.616–620.

Locke, J.C.W. et al., 2006. Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Molecular systems biology*, 2.

Locke, J.C.W. et al., 2005. Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular systems biology*, 1(1), pp.E1–E9.

Ma, H. et al., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue), pp.D258–61.

Maere, S., Heymans, K. & Kuiper, M., 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16), pp.3448–3449.

Makino, S. et al., 2002. The APRR1/TOC1 quintet implicated in circadian rhythms of Arabidopsis thaliana: I. Characterization with APRR1-overexpressing plants. *Plant and Cell Physiology*, 43(1), pp.58–69.

Martin-Tryon, E.L., Kreps, J.A. & Harmer, S.L., 2006. GIGANTEA Acts in Blue Light Signaling and Has Biochemically Separable Roles in Circadian Clock and Flowering Time Regulation. *PLANT PHYSIOLOGY*, 143(1), pp.473–486.

Mas, P. et al., 2003. Targeted degradation of TOC1 by ZTL modulates circadian function in Arabidopsis thaliana. *Nature*, 426(6966), pp.567–570.

Matsuo, T. et al., 2003. Control Mechanism of the Circadian Clock for Timing of Cell Division in Vivo. *Science*, 302(5643), pp.255–235.

McClung, C.R., 2006. Plant circadian rhythms. *The Plant cell*, 18(4), pp.792–803.

McClung, C.R., Salomé, P.A. & Michael, T.P., 2002. The Arabidopsis circadian system. *The Arabidopsis book / American Society of Plant Biologists*, 1, p.e0044.

Michael, T.P., 2003. Enhancer Trapping Reveals Widespread Circadian Clock Transcriptional Control in Arabidopsis. *PLANT PHYSIOLOGY*, 132(2), pp.629–639.

Michaelis, L. & Menten, M.L., 1913. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369), p.352.

Millar, A.J. & Kay, S.A., 1991. Circadian Control of cab Gene Transcription and mRNA Accumulation in Arabidopsis. *The Plant cell*, 3(5), pp.541–550.

Millar, A.J. et al., 1992. A novel circadian phenotype based on firefly luciferase expression in transgenic plants. *The Plant cell*, 4(9), pp.1075–1087.

Millar, A.J. et al., 1995. Circadian clock mutants in Arabidopsis identified by luciferase imaging. *Science*, 267(5201), pp.1161–1163.

Mills, J.N., Minors, D.S. & Waterhouse, J.M., 1974. The circadian rhythms of human subjects without timepieces or indication of the alternation of day and night. *The Journal of physiology*, 240(3), pp.567–594.

Miyama, M. & Tada, Y., Transcriptional and physiological study of the response of Burma mangrove (Bruguiera gymnorhiza) to salt and osmotic stress. *Plant Molecular Biology*, 68(1-2), pp.119–129.

Mizoguchi, T. et al., 2002. LHY and CCA1 Are Partially Redundant Genes Required to Maintain Circadian Rhythms in Arabidopsis. *Developmental Cell*, 2(5), pp.629–641.

Moore et al., 2014. Online Period Estimation and Determination of Rhythmicity in Circadian Data, using BioDare data infrastructure. In press.

Nakajima, M. et al., 2005. Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. *Science*, 308(5720), pp.414–415.

Nakamichi, N., 2005. PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, Together Play Essential Roles Close to the Circadian Clock of Arabidopsis thaliana. *Plant and Cell Physiology*, 46(5), pp.686–698.

Noble, D., 2006. The music of life biology beyond the genome. *public.eblib.com*. Available at: http://public.eblib.com/EBLPublic/PublicView.do?ptiID=431157 [Accessed January 7, 2014].

Ouyang, Y. et al., 1998. Resonating circadian clocks enhance fitness in cyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15), pp.8660–8664.

O'Neill, J.S. et al., 2011. Circadian rhythms persist without transcription in a eukaryote. *Nature*, 469(7331), pp.554–558.

Penfold, C.A. & Wild, D.L., 2011. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6), pp.857–870.

Plautz, J.D. et al., 1997. Quantitative Analysis of Drosophila period Gene Transcription in Living Animals. *Journal of Biological Rhythms*, 12(3), pp.204–217.

Pokhilko, A. et al., 2010. Data assimilation constrains new connectionsand components in a complex, eukaryotic circadian clock model. *Molecular systems biology*, 6, pp.1–10.

Pokhilko, A. et al., 2012. The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Molecular systems biology*, 8, pp.1–13.

Portolés, S. et al., 2010. The Functional Interplay between Protein Kinase CK2 and CCA1 Transcriptional Activity Is Essential for Clock Temperature Compensation in Arabidopsis. *PLoS Genetics*, 6(11), p.e1001201.

Ptitsyn, A., 2008. Comprehensive analysis of circadian periodic pattern in plant transcriptome. *BMC Bioinformatics*, 9(Suppl 9), p.S18.

Queitsch, C. et al., Heat Shock Protein 101 Plays a Crucial Role in Thermotolerance in Arabidopsis. *plantcell.org*.

Reed, J.W. et al., 1994. Phytochrome A and Phytochrome B Have Overlapping but Distinct Functions in Arabidopsis Development. *PLANT PHYSIOLOGY*, 104(4), pp.1139–1149.

Rhee, S.Y., 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, 31(1), pp.224–228.

Richardson, K. et al., 1998. T-DNA tagging of a flowering-time gene and improved gene transfer by in planta transformation of Arabidopsis. *Australian Journal of Plant Physiology*, 25(1), p.125.

Ruoff, P., Loros, J.J. & Dunlap, J.C., 2005. The relationship between FRQ-protein stability and temperature compensation in the Neurospora circadian clock. *Proceedings of the National Academy of Sciences*, 102(49), pp.17681–17686.

Rusak, B. & Zucker, I., 1975. Biological rhythms and animal behavior. *Annual Review of Psychology*, 26(1), pp.137–171.

Rutherford, A.W., Govindjee & Inoue, Y., 1984. Charge accumulation and photochemistry in leaves studied by thermoluminescence and delayed light emission. *Proceedings of the National Academy of Sciences of the United States of America*, 81(4), pp.1107–1111.

Salome, P.A., Penfield, S. & Hall, A., 2005. PSEUDO-RESPONSE REGULATOR 7 and 9 Are Partially Redundant Genes Essential for the Temperature Responsiveness of the Arabidopsis Circadian Clock. *THE PLANT CELL ONLINE*, 17(3), pp.791–803.

Salome, P.A., Weigel, D. & McClung, C.R., 2010. The Role of the Arabidopsis Morning Loop Components CCA1, LHY, PRR7, and PRR9 in Temperature Compensation. *THE PLANT CELL ONLINE*, 22(11), pp.3650–3661.

Savage, N.S. et al., 2008. A Mutual Support Mechanism through Intercellular Movement of CAPRICE and GLABRA3 Can Pattern the Arabidopsis Root Epidermis. *PLoS Biology*, 6(9), p.e235.

Schaffer, R. et al., 1998. The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering. *Cell*, 93(7), pp.1219–1229.

Scholl, R.L., May, S.T. & Ware, D.H., 2000. Seed and molecular resources for Arabidopsis. *PLANT PHYSIOLOGY*, 124(4), pp.1477–1480.

Shannon, P., 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11), pp.2498–2504.

Sideaway-Lee et al., 2014. Arabidopsis transcriptomic responses to ambient temperature are effected predominantly through regulation of transcription rates. In press.

Snoep, J.L. & Westerhoff, H.V., 2005. From isolation to integration, a systems biology approach for building the Silicon Cell. In *link.springer.com*. Topics in Current Genetics. Berlin/Heidelberg: Springer-Verlag, pp. 13–30.

Somers, D.E. et al., The short-period mutant, toc1-1, alters circadian clock regulation of multiple outputs throughout development in Arabidopsis thaliana. *dev.biologists.org*.

Southern, M.M., Brown, P.E. & Hall, A., 2006. Luciferases as reporter genes. *Methods in molecular biology (Clifton, N.J.)*, 323, pp.293–305.

Thimm, O. et al., 2004. mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6), pp.914–939.

Tomita, J., 2005. No Transcription-Translation Feedback in Circadian Rhythm of KaiC Phosphorylation. *Science*, 307(5707), pp.251–254.

Tseng, Y.-Y. et al., 2012. Comprehensive Modelling of the Neurospora Circadian Clock and Its Temperature Compensation. *PLoS Computational Biology*, 8(3), p.e1002437.

Wang, Z.Y. et al., 1997. A Myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene. *The Plant cell*, 9(4), pp.491–507.

Wenden, B. et al., 2011. Light inputs shape the Arabidopsis circadian system. *The Plant Journal*, 66(3), pp.480–491.

Werhli, A.V., Grzegorczyk, M. & Husmeier, D., 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20), pp.2523–2531.

Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis* 2nd ed, Springer Publishing Company, Incorporated.

Windram, O. et al., 2012. Arabidopsis Defense against Botrytis cinerea: Chronology and Regulation Deciphered by High-Resolution Temporal Transcriptomic Analysis. *The Plant cell*, 24(9), pp.3530–3557.

Wold, S., Esbensen, K. & Geladi, P., 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), pp.37–52.

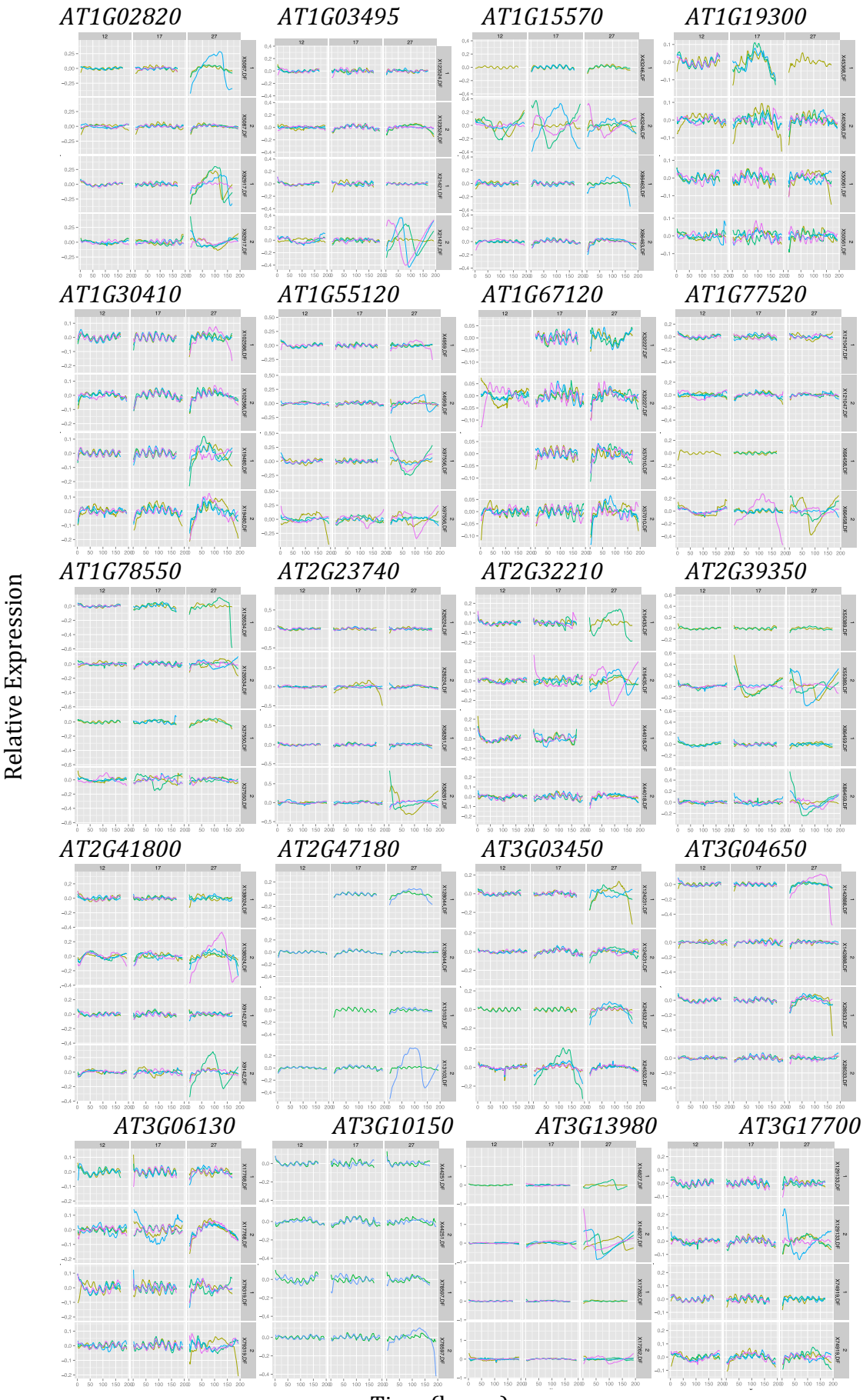
Wu, C. et al., 2005. gcrma: Background Adjustment Using Sequence Information. *R package version*, 2100.

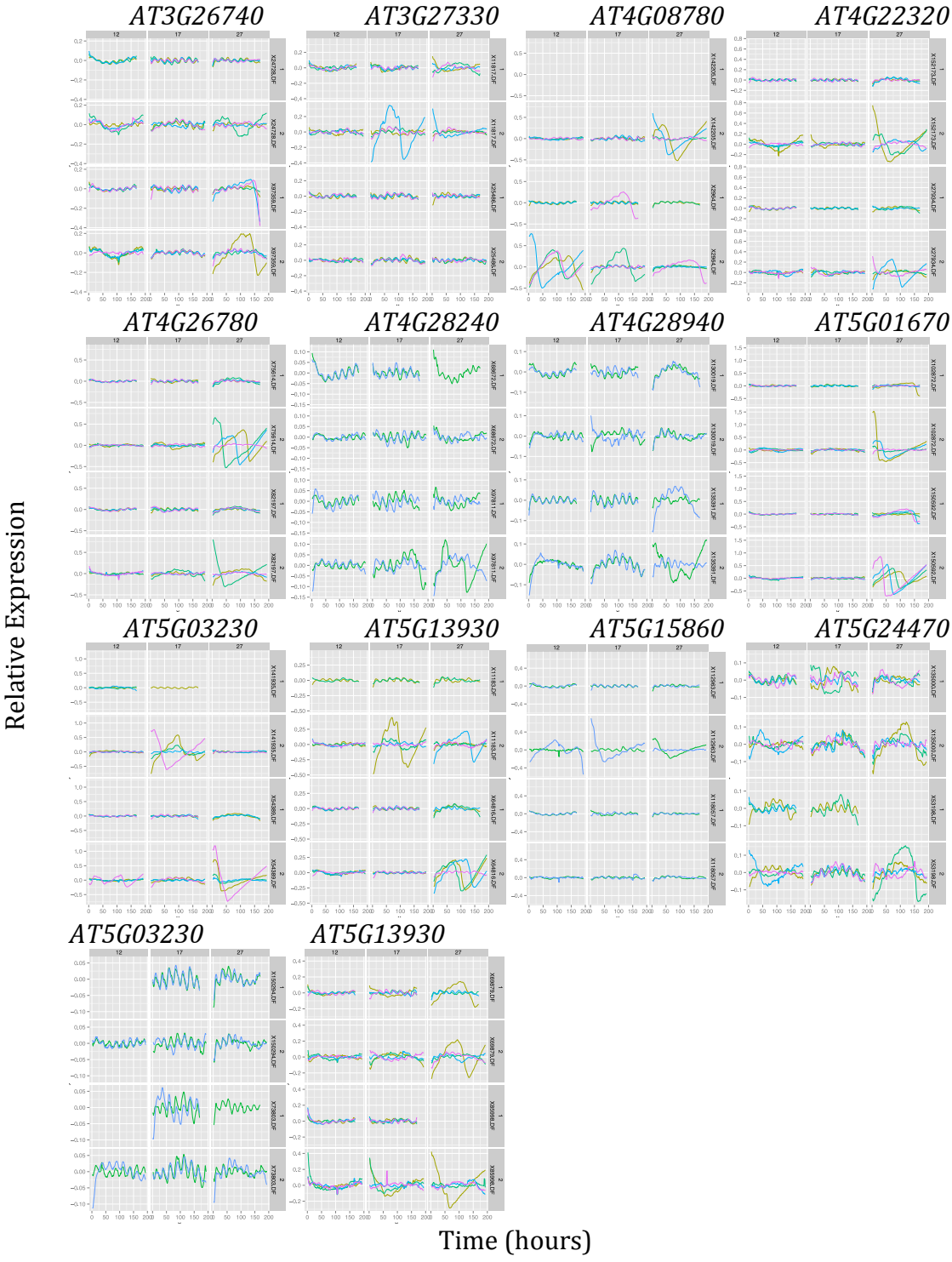
Yu, X. et al., 2010. The Cryptochrome Blue Light Receptors. *The Arabidopsis book / American Society of Plant Biologists*, 8, p.e0135.

Zou, C., Denby, K.J. & Feng, J., 2009. Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics*, 10, p.122.

Zwicker, D., Lubensky, D.K. & Wolde, ten, P.R., 2010. Robust circadian clocks from coupled protein-modification and transcription–translation cycles. *Proceedings of the National Academy of Sciences*, 107(52), pp.22540–22545.

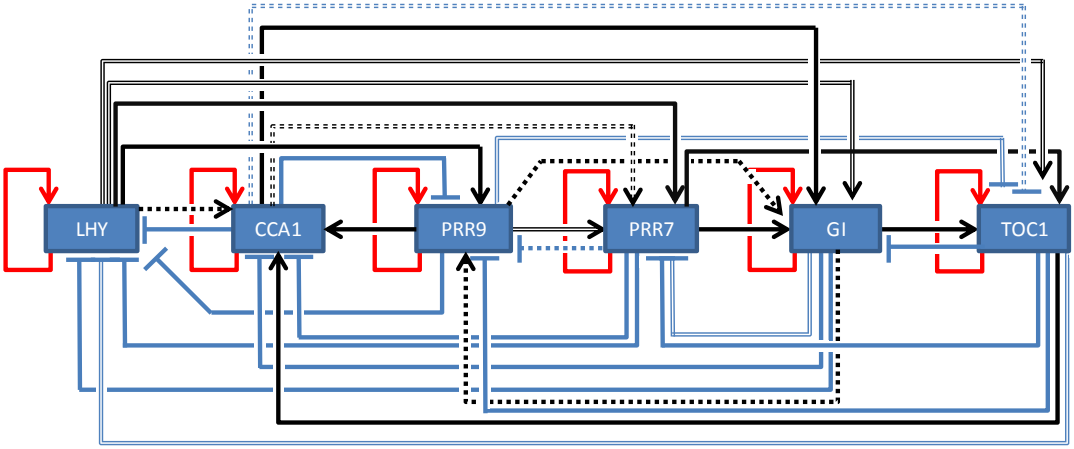
Appendix 1 – Supplemental Data


AT1G02820
AT1G03495
AT1G15570
AT1G19300
AT1G30410
AT1G55120
AT1G67120
AT1G77520
AT1G78550
AT2G23740
AT2G32210
AT2G39350
AT2G41800
AT2G47180
AT3G03450
AT3G04650
AT3G06130
AT3G10150
AT3G13980
AT3G17700
Relative Expression
Time (hours)

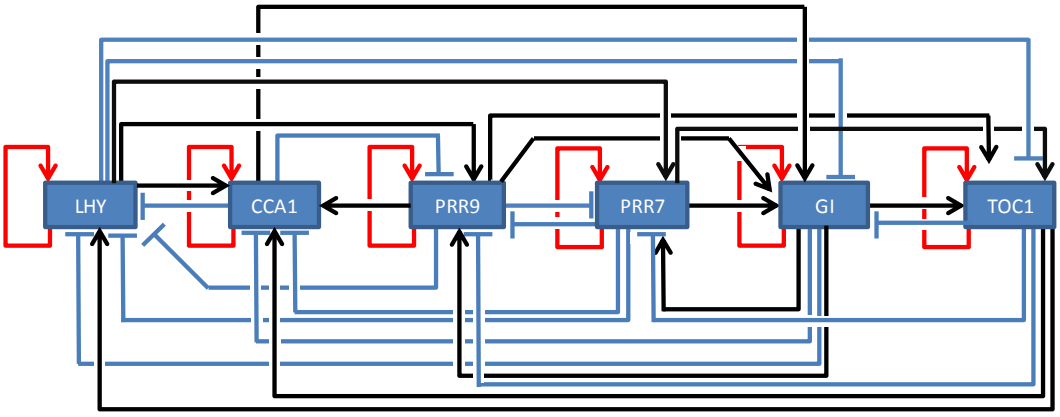


Supplemental Figure 3.1 Delayed fluorescence time course of mutant knockouts. Each gene had two mutants (NASC ID's on the right) and was performed in duplicate. This was done at three temperatures (top) under red/blue light. Each line shows an individual set of plants recorded.

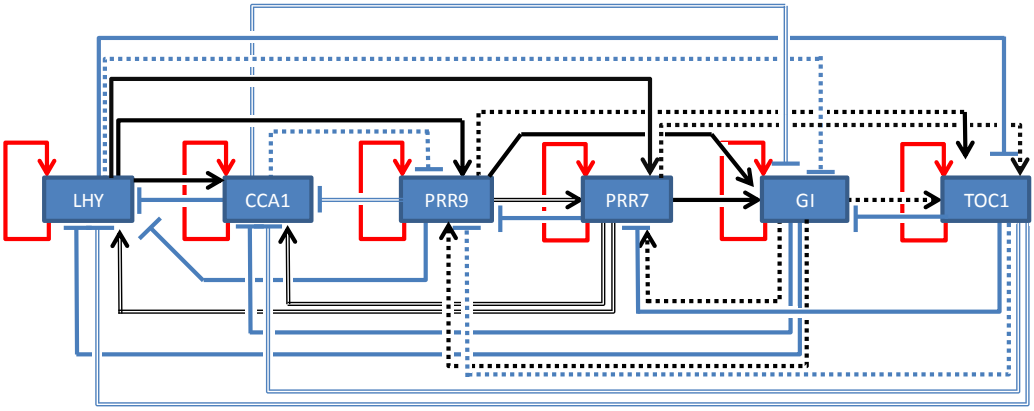
12˚C



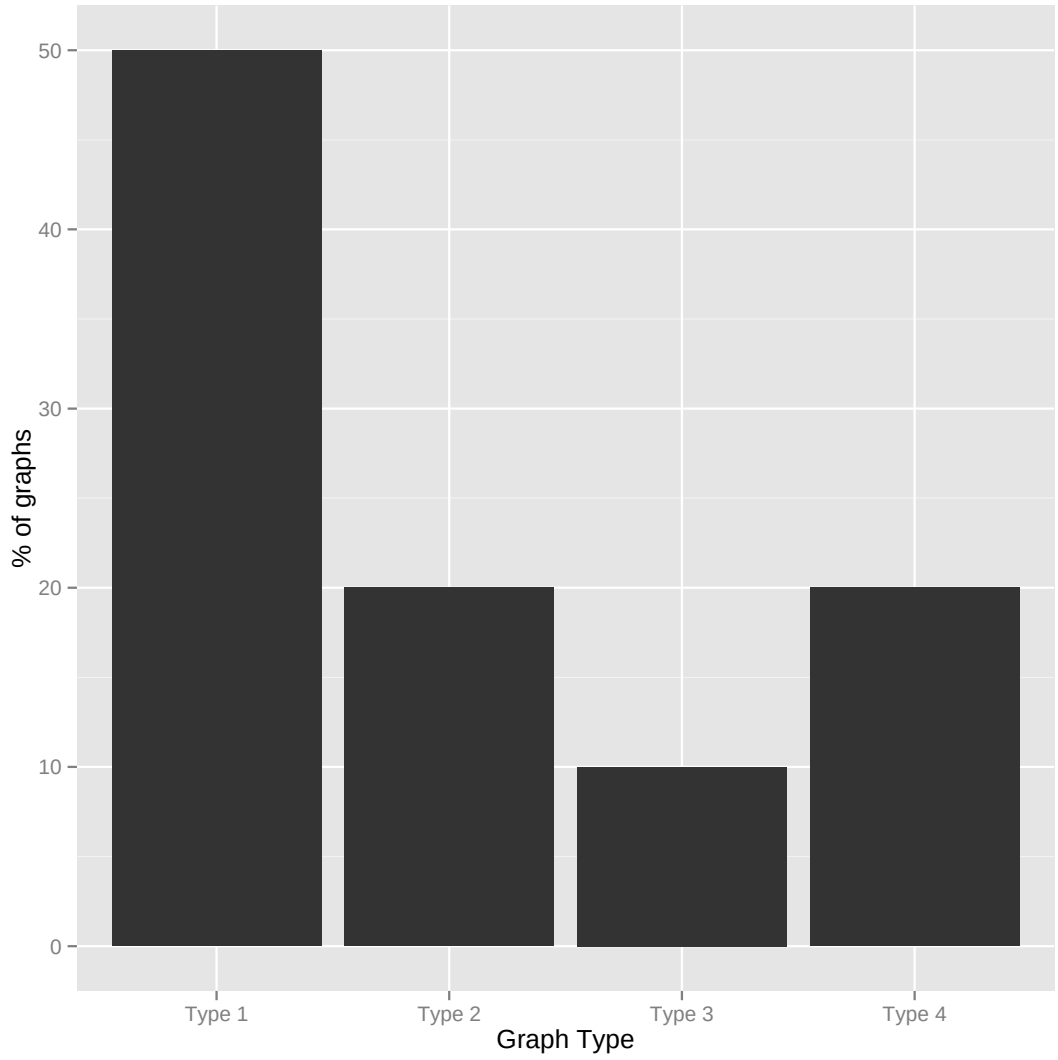
17˚C



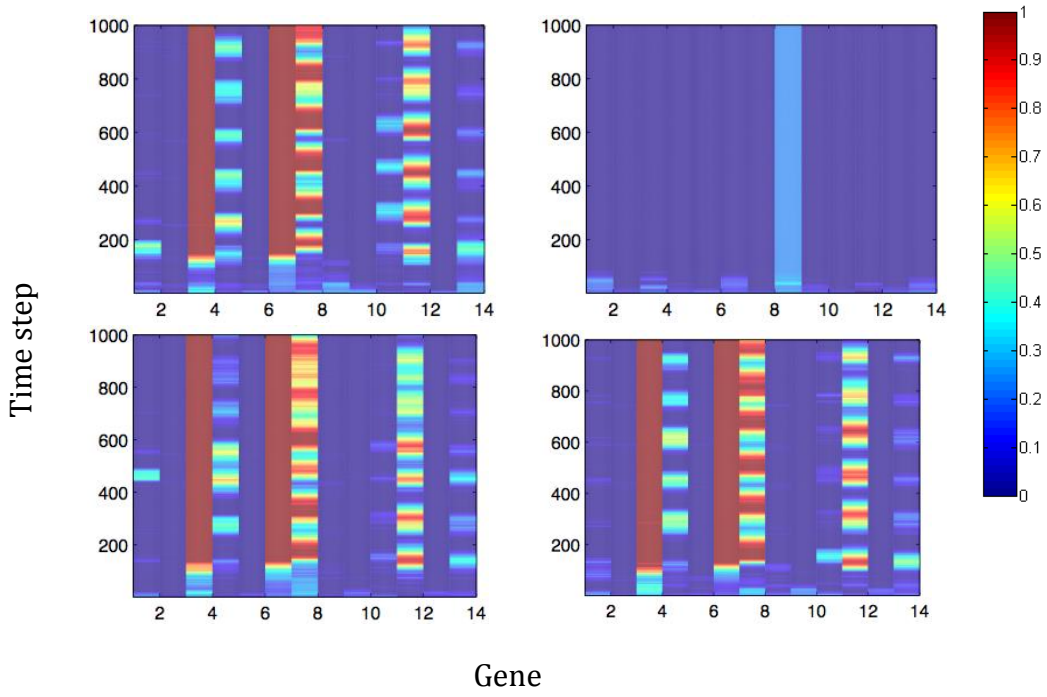
27˚C



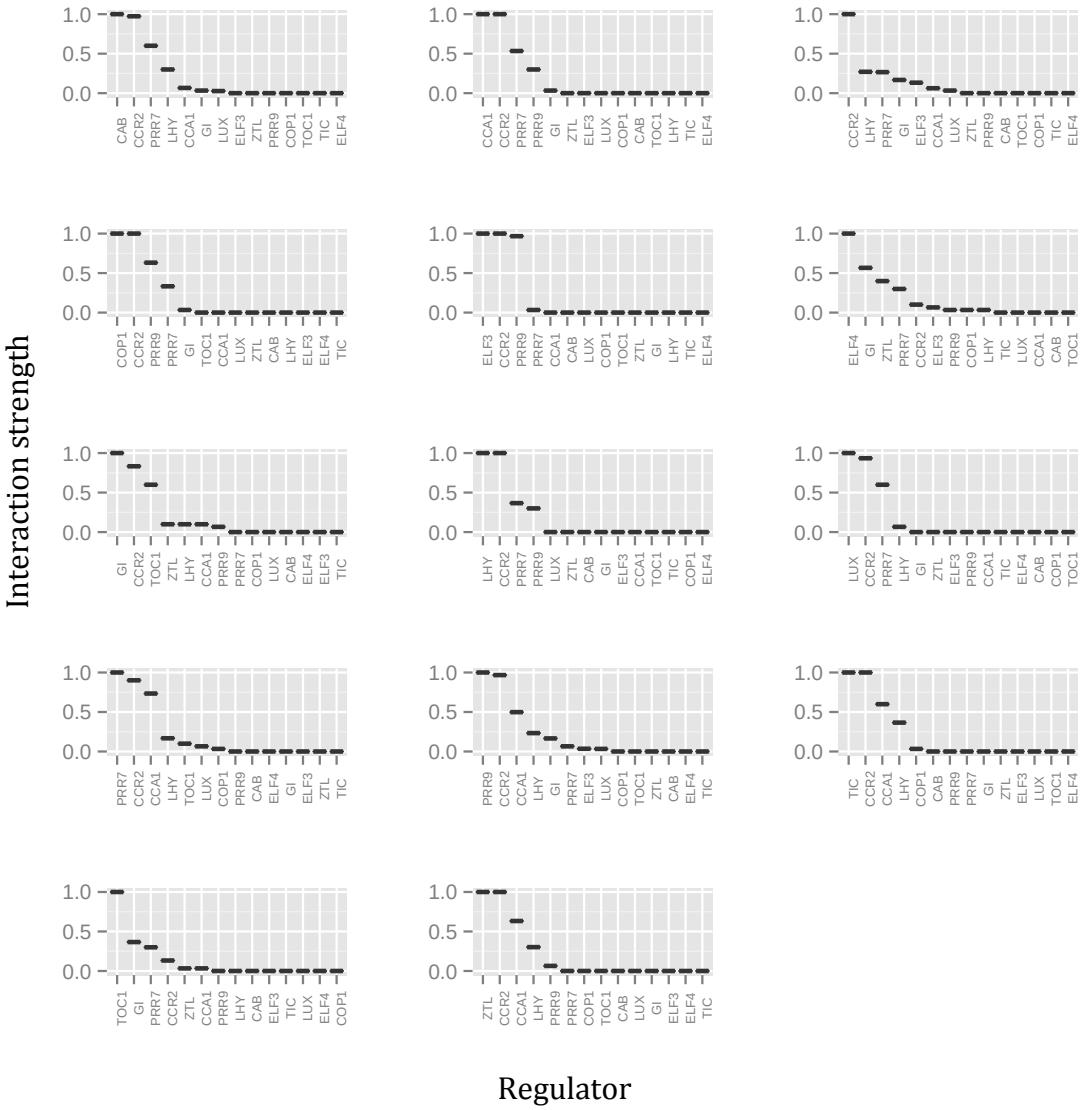
Supplemental Figure 5.1 VBSSM inferred networks for WT luciferase data. Networks for 12˚C and 27˚C are drawn in comparison to 17˚C. Red connections are auto-regulation, black are positive interactions and blue are negative interactions. Dotted lines show a lost interaction and dashes are new interactions.



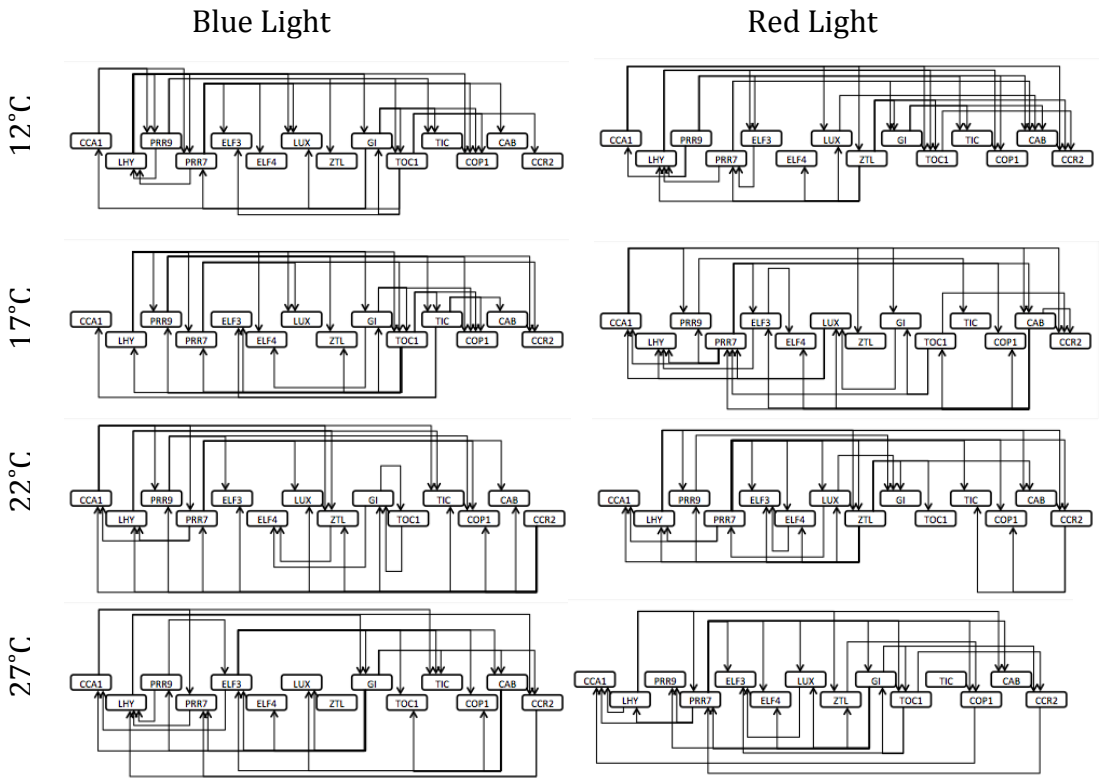
Supplemental Figure 5.2 Percentage of each graph type produced by simulating the matrix produced by VBSSM. VBSSM was run using data generated at 17˚C under blue light conditions, and 14 hidden states.



Supplemental Figure 5.3 Example output of simulating inferred networks using real expression start points. Left column was recovered using the raw matrix, the right column was recovered using the P-value scaled matrix. This was performed using data collected at 17 hours after dawn. Rows represent 2 separate realisations of the simulation (but conserved across the row).



Supplemental Figure 6.1 Connection strengths for each gene in order of strengths. Network inference on the 14 gene network at 22˚C BL, networks are the average of 30 independent runs of CSI using differently combined data. First regulator is self regulation in all cases.



Supplemental Figure 6.2 Visualisation of the inferred network of 14 components. The top 28 connections of the network created by CSI under each condition, using data from the genes present in Pokhilko 2012 plus TIC, CAB2 and CCR2.

Supplemental Table 3.1 Up-regulated genes in response to increased temperature. List of genes that were in clusters whose average expression showed a 2-fold increase over the temperature range. Values are log2 fold change values compared to 17˚C. Table saved as Supplemental Table 3.1.xlsx.

Supplemental Table 3.2 Down-regulated genes in response to increased temperature. List of genes that were in clusters whose average expression showed a 2-fold decrease over the temperature range. Values are log2 fold change values compared to 17˚C. Table saved as Supplemental Table 3.2.xlsx.

Supplemental Table 3.3 Genes differentially expressed in gi- mutant compared to wild type. Lists are split into each individual temperature and whether they were up regulated in the mutant or down regulated. Table saved as Supplemental Table 3.3.xlsx.

Supplemental Table 3.4 List of genes differentially expressed with temperature and in the gi- mutant. Values are the raw reading outputted by gcRMA. Table saved as Supplemental Table 3.4.xlsx.

Supplemental Table 3.5 T-test results on spectral resampling period scores. Highlighted cells are significant at the 5% level, rows with red text only had 1 recorded. Dashes show where too few well had measurements for an accurate t-test. Table saved as Supplemental Table 3.5.xlsx.

Supplemental Table 4.1 Cluster membership of luciferase expression. Luciferase expression for each condition was clustered independently using SplineCluster.

Supplemental Table 4.2 Cluster membership of averaged luciferase expression. Luciferase expression for each condition was clustered independently using SplineCluster.

	12C									17C									22C									27C								
	RL			BL			RBL			RL			BL			RBL			RL			BL			RBL			RL			BL			RBL		
Gene ID	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL	LDLL	LD	LL
ARR4	6	2	7	4	1	4	2	2	2	6	1	1	1	4	2	1	1	1	2	2	3	1	4	6	4	2	4	5	2	3	3	2	2	4	1	1
ARR9	4	1	3	3	1	3	6	4	3	3	2	4	4	1	1	4	3	2	4	2	2	3	2	5	3	1	2	1	3	1	6	1	7	5	4	4
bHLH16	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	2	3	1	4	2	1	4	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	3	4	4	1	5
CAB	5	2	2	4	1	4	7	1	4	6	2	3	2	4	2	2	4	2	2	2	2	4	4	4	2	1	1	2	2	1	6	3	7	2	2	5
CAT3	5	1	1	6	2	5	1	1	1	6	1	1	3	4	3	1	1	1	2	2	3	4	4	1	4	2	4	5	4	3	1	2	3	3	2	5
CBF1	2	2	5	1	1	1	4	2	6	1	2	6	5	4	4	6	1	4	5	2	5	6	3	2	5	3	5	5	2	4	1	2	3	2	2	3
CBF2	2	3	5	1	1	1	4	2	6	1	2	6	4	1	4	1	1	1	5	2	5	5	3	3	5	3	5	6	4	4	2	3	1	2	2	3
CBF3	3	2	4	4	1	4	7	1	4	6	1	3	4	1	1	1	1	1	2	2	2	3	2	4	2	1	1	4	4	2	6	1	7	2	1	3
CCA1	5	1	2	4	1	4	7	1	4	4	3	2	2	3	2	2	4	3	1	1	1	2	1	6	1	1	3	2	2	1	5	1	6	6	4	2
CCR2	2	3	5	1	1	1	4	2	6	1	2	6	5	2	4	6	6	4	5	2	5	6	3	2	5	3	5	6	2	4	3	2	1	1	3	3
CO	6	2	7	6	1	5	2	2	2	6	2	3	1	4	3	1	1	1	2	2	3	1	4	6	4	2	4	6	4	1	4	3	5	4	1	1
COP1	5	1	2	4	1	4	2	1	2	4	3	3	1	4	2	2	4	3	2	2	2	1	4	6	1	1	3	3	1	3	4	3	4	4	1	1
COR15A	1	2	6	1	1	1	4	2	6	2	2	5	5	2	4	6	1	4	5	2	5	5	3	3	5	3	5	6	4	4	2	3	1	2	2	5
CRY1	5	2	1	4	1	4	2	1	2	6	2	3	1	4	2	1	1	1	2	2	2	1	4	6	1	1	3	4	2	1	4	3	4	4	1	1
CRY2	6	2	7	5	1	5	1	1	1	6	1	1	3	4	3	1	1	1	2	2	3	7	4	1	4	2	4	7	4	5	1	2	3	3	2	5
ELF3	5	1	2	4	1	4	2	1	2	4	3	3	1	3	2	2	4	3	1	1	1	1	4	6	1	1	3	2	4	1	4	3	5	4	1	1
ELF4	2	3	5	1	1	1	3	3	5	6	2	3	5	2	4	6	6	4	5	2	4	6	3	2	5	3	5	5	4	4	1	3	3	2	3	3
FAR1	5	1	1	6	2	5	1	1	1	6	2	3	3	4	3	1	1	1	2	2	3	7	4	1	4	2	4	5	2	3	1	2	3	3	2	5
FHY1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	2	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	2	5
FHY1Like	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	5	2	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2	2	5
GAI	NA	NA	NA	NA	NA	NA	NA	NA	NA	6	2	1	3	4	3	5	2	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	3	3	2	2	5
GI	2	3	5	1	1	1	4	2	4	1	2	6	5	2	4	6	6	4	5	2	5	5	3	3	5	3	5	6	2	4	3	3	1	2	3	3
HFR1	6	2	7	4	1	4	7	1	4	6	1	3	4	1	3	1	4	1	2	2	2	4	3	4	2	1	1	3	1	3	6	2	2	2	2	5
HY5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	5	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6	4	3
HYH	5	1	3	4	1	4	7	1	2	4	3	3	1	3	2	2	4	2	1	1	1	2	1	6	1	1	3	5	2	3	4	1	2	6	4	1
ICE1	6	2	7	6	1	5	2	2	2	6	1	1	1	4	3	1	1	1	2	2	3	1	4	6	4	2	4	6	4	5	1	2	2	4	1	1
LHY	5	1	2	4	1	4	7	1	4	4	3	2	2	3	2	2	4	3	1	1	1	2	1	6	1	1	3	2	2	1	5	1	6	6	4	2
LUX	2	2	5	1	1	1	NA	NA	NA	1	2	6	5	2	4	6	6	4	5	2	4	6	3	2	5	3	5	6	4	4	3	2	1	1	3	3
PHYA	6	2	7	5	1	5	1	1	1	6	1	1	3	4	3	1	1	1	2	2	3	4	3	1	4	2	4	7	4	5	1	2	3	2	2	5
PHYB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	2	1
PHYC	6	2	7	6	2	5	1	1	1	6	1	1	3	4	3	1	1	1	2	2	3	7	4	1	4	2	4	7	4	5	1	3	3	4	1	1
PHYD	NA	NA	NA	NA	NA	NA	NA	NA	NA	6	1	1	3	4	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	3	4	NA	NA	NA
PIF3	6	2	7	6	1	5	1	2	6	6	2	1	3	4	3	1	1	1	2	2	3	7	4	1	4	2	4	7	4	5	1	3	3	3	2	5
PIF4	2	2	5	2	1	2	5	2	4	6	2	3	4	1	1	4	2	2	2	2	2	4	3	4	2	3	1	4	2	2	6	3	1	2	2	5
PIF5	5	2	1	4	1	4	7	1	4	6	1	3	4	1	1	4	2	2	3	2	2	4	3	4	2	1	1	4	2	2	6	2	7	2	2	5
PRR3	6	2	5	6	2	5	1	2	6	1	2	6	5	4	4	6	1	4	5	2	4	6	3	1	5	3	5	5	2	3	3	2	2	2	3	3
PRR7	2	3	5	2	1	2	4	2	4	6	2	3	4	1	1	NA	NA	NA	2	2	5	4	3	3	2	3	1	4	2	2	2	3	1	NA	NA	NA
PRR9	4	1	3	3	1	3	6	4	3	3	2	4	4	1	1	4	3	2	4	2	2	3	2	5	3	1	2	1	3	1	6	1	7	5	4	4
RGL3	NA	NA	NA	NA	NA	NA	NA	NA	NA	6	2	3	4	4	3	5	2	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2	3	1	2	2	5
SPA2	6	2	7	6	2	5	1	2	1	6	1	1	3	4	3	5	2	2	2	2	3	7	4	1	4	2	4	7	4	5	1	2	3	3	2	5
SPA3	5	1	2	4	1	4	2	1	2	4	3	3	1	3	2	2	4	3	1	1	1	1	4	6	1	1	3	2	2	1	4	3	5	4	1	1
SPA4	5	1	1	4	1	4	7	1	2	6	2	3	1	3	2	2	4	3	2	2	3	1	4	6	1	1	3	5	2	3	4	3	2	4	1	1
TIC	6	2	7	6	2	5	1	1	1	6	1	1	3	4	3	1	1	1	3	2	2	4	3	1	4	2	4	7	4	5	1	2	3	3	2	5
TOC1	2	3	5	1	1	1	4	2	6	1	2	6	5	2	4	6	6	4	5	2	4	6	3	2	5	3	5	6	4	4	3	2	1	1	3	3
ZTL	5	2	1	6	2	5	2	2	2	6	1	3	3	4	3	1	1	1	3	2	2	7	4	1	4	2	4	4	2	5	1	3	3	4	1	1

Supplemental Table 6.1 Connection strengths of inferred 7-gene networks. Raw CSI output for each set of conditions. Within a square, the top left value is the strength at 12˚C, top right 17˚C, bottom left 22˚C and bottom right 27˚C. The top Matrix was generated in blue light and the bottom one in red light.

BL	LHY/CCA1		PRR7/9		ELF4		GI		TOC1	
LHY/CCA1			0.99999	0.65398	9.42E-06	0.03334	0.29716	0.23322	0.23509	0.96667
			0.9	0.60009	1.71E-41	0.40126	0.96667	0.13347	0.06667	0.86636
PRR7/9	0.66148	0.86667			2.90E-17	0.13303	0.30061	0.10269	0.96665	0.86666
	0.66667	0.8			0.00104	0.40918	0.83357	0.66687	0.4175	0.1998
ELF4	0.30834	0.83038	0.71705	0.10001			0.21569	0.80518	0.06726	0.03333
	0.58113	0.79843	0.50022	0.13336			0.53333	0.63325	0.13333	0.23341
GI	0.54715	0.93523	0.26615	0.33333	7.35E-26	0.06667			1	0.46669
	0.50155	0.4332	0.36667	0.56698	6.62E-14	0.3			0.56667	0.53345
TOC1	0.09759	0.93119	0.83574	0.44608	1.10E-31	8.61E-34	0.90241	0.55392		
	0.31308	0.77479	0.36657	0.13334	0.03342	0.09978	0.83517	0.63355		

RL	LHY/CCA1		PRR7/9		ELF4		GI		TOC1	
LHY/CCA1			0.72434	0.53328	0.37528	0.13276	0.16667	0.30002	1.6E-07	0.15458
			0.79987	0.96667	1.7E-31	3.8E-84	0.46845	3.1E-34	0.16684	0.03333
PRR7/9	0.60071	0.93335			0.43327	0.29212	0.06667	0.40099	0.70054	0.06676
	0.75336	0.83337			0.03329	0.06667	0.93333	1.1E-07	0.90005	0.19985
ELF4	0.13336	0.6665	0.60006	0.9			0.10001	0.06672	0.6334	0.23344
	0.9002	0.56979	0.93333	1			3.4E-07	1.2E-06	0.16647	0.19669
GI	0.19773	0.23531	0.71546	0.58664	0.4887	0.27998			0.06394	0.37636
	0.40149	0.43852	0.26674	0.5692	0.18511	0.13276			0.80036	0.4003
TOC1	0.53487	0.33817	0.46513	0.32849	0.84249	0.86665	0.03333	0.1		
	0.68923	0.46608	0.69997	0.83421	0.11095	0.03337	0.26642	6E-10		

Supplemental Table 6.2 Connection strengths of inferred 14-gene networks. Raw CSI output for each set of conditions. Within a square, the top left value is the strength at 12˚C, top right 17˚C, bottom left 22˚C and bottom right 27˚C. The top Matrix was generated in blue light and the bottom one in red light.

BL	CAB		CCA1		CCR2		COP1		ELF3		ELF4		GI		LHY		LUX		PRR7		PRR9		TIC		TOC1		ZTL	
CAB			0	0	0	0	0	0	0	0.03	0.03	0.03	0.13	0.17	0	0.07	0.1	0	0.83	0.2	0.17	0.17	0	0.59	0.07	0.27	0	0
			0.07	0.2	0.97	0	0	0	0	0.6	0	0	0.03	0.8	0.3	0.07	0.03	0	0.6	0.07	0	0	0	0	0	0	0	0
CCA1	0	0			0	0	0	0	0	0	0	0.03	0.3	0.1	0.03	0	0	0.16	0.8	0.16	0.13	0.2	0	0.5	0.13	0.17	0	0
	0	0.23			1	0	0	0.03	0	0.68	0	0	0.03	0.23	0	0.2	0	0	0.53	0.07	0.3	0.16	0	0.13	0	0.03	0	0.03
CCR2	0	0	0.23	0.03			0.03	0	0	0	0	0.03	0.13	0.03	0.23	0.27	0	0	0.17	0.46	0.03	0.34	0	0	0.7	0.1	0	0.13
	0	0	0.06	0			0	0.1	0.13	0	0	0	0.17	0.4	0.27	0.53	0.03	0	0.27	0.17	0	0	0	0.03	0	0	0	0
COP1	0.02	0	0	0	0	0			0	0	0	0	0.28	0.56	0.3	0	0.07	0.03	0.69	0.01	0.03	0.37	0	0.37	0.41	0.6	0	0
	0	0.27	0	0.2	1	0			0	0.57	0	0	0.03	0.13	0	0.03	0	0	0.33	0.03	0.63	0	0	0	0	0.07	0	0.03
ELF3	0	0	0	0.03	0	0	0	0			0	0	0.18	0.3	0.13	0	0	0	0.75	0.03	0.07	0.17	0	0.63	0.73	0.43	0	0
	0	0.47	0	0.03	1	0	0	0.03			0	0	0	0.3	0	0.1	0	0	0.03	0	0.97	0.4	0	0	0	0	0	0
ELF4	0.06	0	0.03	0.07	0	0	0.03	0	0	0.03			0.17	0.53	0.27	0.83	0	0	0.8	0.13	0.1	0.07	0	0	0.03	0	0.03	0
	0	0	0	0	0.1	0.13	0.03	0.03	0.07	0.13			0.57	0.53	0.03	0.17	0	0	0.3	0.07	0.03	0.07	0	0	0	0.07	0.4	0
GI	0.13	0.03	0.1	0.03	0	0	0	0	0.03	0	0	0.03			0.43	0.67	0	0.03	0	0.03	0.07	0.13	0	0	0.87	0.43	0.03	0.07
	0	0.17	0.1	0	0.83	0	0	0.03	0	0.5	0	0			0.1	0.6	0	0	0	0.1	0.07	0.1	0	0	0.6	0	0.1	0.03
LHY	0	0	0	0	0	0	0	0	0.03	0	0	0.03	0.1	0			0.03	0.17	0.9	0.13	0.33	0.24	0	0.22	0	0.38	0	0
	0	0.04	0	0	1	0.52	0	0.03	0	0	0	0	0	0.07			0	0	0.37	0.27	0.3	0.6	0	0	0	0	0	0
LUX	0.03	0	0.17	0	0	0	0.03	0	0	0.03	0	0.03	0.43	0.1	0.3	0.83			0.43	0.57	0.2	0.1	0	0	0	0	0	0
	0	0.3	0	0	0.93	0.17	0	0.03	0	0.07	0	0.03	0	0.33	0.07	0.17			0.6	0.17	0	0.1	0	0	0	0.03	0	0
PRR7	0.07	0	0.3	0.17	0	0	0	0.03	0	0	0	0.13	0.06	0.03	0.7	0.37	0.06	0.1			0.04	0	0	0.03	0.63	0.67	0	0.33
	0	0.08	0.73	0.29	0.9	0.33	0.03	0	0	0	0	0	0	0.35	0.17	0.07	0.07	0			0	0	0	0	0.1	0.02	0	0.03
PRR9	0.2	0	0.5	0.2	0	0	0	0	0	0	0	0.07	0.13	0.03	0.3	0.73	0.07	0.16	0.23	0			0	0.24	0.03	0.37	0	0
	0	0.12	0.5	0.09	0.97	0	0	0	0.04	0.22	0	0	0.17	0.91	0.23	0	0.03	0	0.07	0			0	0	0	0	0	0.03
TIC	0.03	0	0.02	0.24	0	0	0.03	0	0	0.03	0	0.03	0.28	0.23	0.13	0.03	0.03	0.1	0.11	0	0.73	0.37			0.01	0.5	0.03	0
	0	0	0.6	0.26	1	0	0.03	0	0	0.93	0	0	0	0.61	0.37	0	0	0	0	0.1	0	0.03			0	0	0	0
TOC1	0	0	0	0	0	0	0.2	0	0	0.03	0	0	0.4	0.43	0	0.57	0	0	0	0.43	0.8	0	0	0			0	0.07
	0	0.23	0.03	0.13	0.13	0	0	0	0	0.27	0	0.17	0.37	0.2	0	0.1	0	0	0.3	0.1	0	0.07	0	0			0.03	0
ZTL	0	0	0.23	0	0	0.03	0.13	0	0.03	0.03	0	0.03	0.03	0.13	0.1	0.03	0.07	0.03	0.07	0.27	0.47	0.5	0	0.03	0.13	0.5		
	0	0	0.63	0.09	1	0.07	0	0.03	0	0.2	0	0	0	0.22	0.3	0.13	0	0.09	0	0	0.06	0	0	0.03	0	0		

RL	CAB		CCA1		CCR2		COP1		ELF3		ELF4		GI		LHY		LUX		PRR7		PRR9		TIC		TOC1		ZTL	
CAB			0	0.83	0	0	0	0	0	0.17	0	0	0.33	0	0	0	0.33	0	0.33	0.83	0.5	0.17	0	0	0.17	0	0	0
			0.33	0.16	0	0	0	0	0.17	0	0	0.17	0	0.17	0.5	0.51	0	0	0	0.67	0.17	0.17	0	0	0	0	0.83	0.17
CCA1	0	0.17			0	0	0	0	0	0	0	0	0	0	0	0	0.33	0.5	0.33	1	0.5	0	0	0	0.33	0	0.17	0
	0	0			0	0	0	0.33	0	0	0	0	0	0	0	0.33	0	0.16	1	1	0.17	0.17	0	0	0	0	0.83	0
CCR2	0	0.49	0.5	0.34			0	0	0.33	0	0	0	0	0	0	0.16	0	0.17	0	0	0	0	0	0	0.5	0.83	0.34	0
	0	0	0	0			0	0	0	0	0	0	0	0.33	0.67	0	0	0	0.5	0	0.33	0	0	0	0	0.33	0.17	0
COP1	0	1	0	0	0	0			0	0	0.03	0	0	0	0.67	0	0	0	0.13	0.83	0.67	0.17	0	0	0	0	0.17	0
	0	0	0	0	0.67	0			0	0	0	0	0	0.01	0	0.01	0	0	0.67	1	0	0	0	0.01	0	0.01	0	0.33
ELF3	0	0.33	0	0	0	0	0	0			0	0	0.17	0.17	0.33	0	0.17	0.17	0.17	0.83	0.67	0.17	0	0	0	0	0.17	0
	0	0	0.17	0	0	0	0	0			0.5	0	0	0.01	0	0	0	0.67	0.67	0.99	0.17	0	0	0	0	0.33	0.5	0
ELF4	0	0.5	0.17	0.17	0	0	0	0	0.32	0.33			0	0	0.01	0.33	0	0.32	0.01	0.17	0	0	0	0	0	0.18	0.82	0
	0	0	0	0.17	0	0.11	0	0	0	0			0.33	0.39	0	0.17	0	0	0.5	1	0.33	0.17	0	0	0	0	0.5	0
GI	0	0.17	0.17	0.66	0	0	0	0	0.17	0.17	0	0			0	0	0	0.17	0.33	0	0	0.17	0.17	0	0.17	0.66	0.67	0
	0	0	0	0	0	0	0	0	0	0	0	0			0	0	0.61	0	0	0.67	0.39	0	0	0	0	1	0.67	0
LHY	0	0	0	0	0	0	0	0	0.17	0.33	0	0.17	0	0.17			0	0.5	0.33	0.67	0.5	0.17	0	0	0	0	0.67	0
	0	0	0	0	0.17	0	0	0	0	0	0	0	0	0.16			0	0	0.83	0.84	0.17	0.17	0	0	0	0.17	0.83	0
LUX	0	0.44	0.83	0	0	0	0	0	0	0	0	0	0	1	0	0			0	0	0	0.17	0	0.17	0	0	0.83	0.11
	0	0	0	0	0	0	0	0	0	0	0	0	0.33	0.83	0	0			0.67	1	0	0	0	0	0	0	1	0.17
PRR7	0	0.5	0.17	0.26	0	0	0	0	0.33	0.17	0	0	0	0	0	0.07	0	0.5			0	0.17	0	0	0	0.33	0.83	0
	0	0	0	0	0	0.33	0	0	0	0	0	0	0	0.67	0	0.83	1	0			0	0.17	0	0	0	0	0.33	0
PRR9	0.17	0	0.33	0.83	0.17	0	0	0	0.17	0.17	0.17	0	0.17	0	0	0	0.17	0	0	0.67			0.17	0	0.17	0	0.17	0
	0	0	0.16	0	0	0	0	0	0	0	0.17	0	0	0.5	0.67	0	0	0	0.17	1			0	0	0	0	0.83	0.33
TIC	0.17	0	0.17	0.17	0.17	0	0.17	0	0.17	0.18	0.17	0	0.33	0	0.17	0	0.17	0	0.17	0.17	0.33	0.67			0.17	0.16	0.17	0
	0	0	0.34	0	0.5	0.16	0	0.33	0	0	0	0	0	0.01	0.17	0	0	0	0.67	0	0.33	0.17			0	0	0	0
TOC1	0.17	0.76	0.67	0.17	0.17	0.02	0.17	0	0.17	0	0.17	0	0.17	0.31	0.33	0.02	0.17	0	0	0	0.17	0	0.17	0			0.67	0
	0	0	0	0	0	0	0	0	0	0	0	0	0.33	1	0	0	0	0	0.33	1	0	0	0	0			1	0
ZTL	0	0	0.33	0.67	0.17	0	0	0	0	0	0	0	0.17	0.17	0	0.17	0	0	0	0.33	0	0	0	0	0	0		
	0	0	0	0	0.17	0	0	0.17	0	0	0	0	0	0.5	0.83	0.5	0	0	0.83	0.17	0.17	0	0	0	0	0		

Supplemental Table 6.3 Connection strengths of inferred 7-gene networks. Raw CSI output for each set of conditions. Within a square, the top left value is the strength at 12˚C, top right 17˚C, bottom left 22˚C and bottom right 27˚C. The top Matrix was generated in blue light and the bottom one in red light.

BL	LHY/CCA1		PRR7/9		ELF4		GI		TOC1	
LHY/CCA1			-0.99999	-0.65398	9.42E-06	0.033337	-0.29716	-0.23322	0.235088	-0.96667
			-0.9	-0.60009	1.71E-41	0.401264	0.966666	0.133468	-0.06667	0.866362
PRR7/9	0.661476	0.866666			2.90E-17	0.13303	-0.30061	-0.10269	-0.96665	-0.86666
	0.666665	0.800001			0.001039	0.409185	0.833566	-0.66687	-0.4175	-0.1998
ELF4	0.308345	0.830381	0.71705	-0.10001			0.215686	0.80518	-0.06726	-0.03333
	0.581131	0.798435	0.500215	0.133358			0.533333	0.633248	-0.13333	0.233415
GI	0.547147	0.935234	0.266153	0.333333	-7.35E-26	0.066667			-1	-0.46669
	0.501552	0.433202	-0.36667	0.566981	-6.62E-14	-0.3			0.566667	0.533445
TOC1	0.09759	0.93119	0.835743	0.44608	1.10E-31	8.61E-34	0.90241	0.55392		
	-0.31308	-0.77479	-0.36657	0.133339	-0.03342	-0.09978	0.835167	-0.63355		

RL	LHY/CCA1		PRR7/9		ELF4		GI		TOC1	
LHY/CCA1			-0.93333	-1	0.033349	0.366665	-0.03333	0.333335	0.788345	0.066667
			-0.96667	-1	0.833333	-0.06666	1.98E-28	-0.06667	1.26E-40	0.1
PRR7/9	0.866427	0.866665			-0.06667	-0.06667	0.566906	-0.73332	-0.70004	-0.93333
	0.199109	0.699983			-0.99976	-1.55E-05	-0.03332	-0.89964	-0.10024	0.100363
ELF4	-0.66714	0.813802	0.242794	0.251189			0.329228	-0.14565	-0.32717	-0.59651
	-0.58403	0.133329	0.833343	0.83751			0.203095	0.538236	7.66E-24	-1.84E-06
GI	-0.23331	0.933323	0.766676	0.06666	1.87E-05	-0.03334			-1	-0.96667
	-7.68E-26	2.79E-54	0.599961	1	-0.56675	-1.35E-07			-0.43325	-1
TOC1	0.499649	0.659107	0.467265	0.066847	4.27E-19	0.03333	0.533729	0.900198		
	-0.49689	-5.80E-13	0.857692	1	0.133344	7.71E-13	-0.39997	1		